E 341 Block2

THE OPEN UNIVERSITY
Educational Studies : A Third Level Course
Methods of Educational Enquiry Block 2

Research Design

SY 0009257 6

ST. MARY'S
COLLEGE OF EDUCATION
LIBRARY
S/370.78
OPE
33901

ARY'S COLLEGE OF EDUCATION
LIBRARY

1·1

The Open University

Educational Studies: A Third Level Course
Methods of Educational Enquiry Block 2

Research Design

Prepared by Neville Bennett for the Course Team

THE OPEN UNIVERSITY PRESS

33901

THE METHODS OF EDUCATIONAL ENQUIRY COURSE TEAM

Phillip Williams (Chairman)
John Bynner
Barry Commins
Beryl Crooks
Jane Greenwald
Donald Holms
Richard Hooper
Victor Lockwood
Ian Macdonald
Tam Murrell
Anthony Pickup
William Prescott
Joan Whitehead
Carolin Wilson

External consultants
Neville Bennett
Noel Entwistle
Desmond Furneaux
Bryan Moore
Desmond Nuttall
Albert Pilliner
Alan Willmott

Course edited by Phillip Williams and John Bynner

The course team are indebted to the following for their assistance: Philip Vernon, Professor of Educational Psychology at the University of Calgary, for reading and commenting on the course material; Patrick Miller, Department of Psychology, University College of Swansea, for preparing the glossary which appears at the end of each volume; John Murphy for assistance in preparing the statistical block of the course; Maurice Chazan, Reader in Education at the University College of Swansea, and Dr Alan Little, Head of the Reference Division of the Community Relations Commission, for preparing materials for the evaluation and assessment block of the course; Joan Barker Lunn and Douglas Pidgeon, Executive Director of the ita Foundation, for their participation in the initial course planning; and the group at the Dacorum Teachers' Centre, Hertfordshire, who developmentally tested the correspondence texts. The final responsibility for the course nevertheless rests with the course team.

The Open University Press
Walton Hall Bletchley Buckinghamshire

First published 1973
Copyright © 1973 The Open University

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publishers.

Designed by the Media Development Group of The Open University

Printed in Great Britain by
EYRE AND SPOTTISWOODE LIMITED
AT GROSVENOR PRESS PORTSMOUTH

SBN 335 03321 0

This text forms part of a series of blocks that make up the correspondence element of an Open University third level course. The complete list of blocks in the course is given at the end of this text.

For general availability of supporting material referred to in this book please write to the Director of Marketing, The Open University, Walton Hall, Bletchley, Buckinghamshire.

Further information on Open University courses may be obtained from the Admissions Office, The Open University, P.O. Box 48, Bletchley, Buckinghamshire.

General Contents

	Page
Introduction to Block 2	5
Part 1 Hypothesis Formulation	7
Part 2 Sampling	33
Glossary for Block 2	69

Introduction to Block 2

In this block we start to transform some of the ideas discussed in Block 1 into the practical decisions that have to be taken in the research process. We look at the two main questions in designing research: how to formulate the hypotheses that are to be tested and how to sample the population in which the hypotheses are to be investigated. In looking at both these topics we stay within the broad limits of what was called 'survey research' in Block 1. Experiments involving the control of certain variables postulated as the cause of the phenomena under investigation require special design procedures, which we shall be looking at in Block 5.

Our commitment to survey research imposes limitations on the conclusions we draw from our investigations: we can make inferences only about the existence of *relationships* between the variables to which our hypotheses refer and not their *direction*; i.e. we cannot strictly say that changes in one variable *produce* changes in another variable, only that the two variables change *concomitantly*. Despite this scientific weakness of survey research, quite strong explanatory inferences can be drawn from it. Some variables clearly antecede others in a causal sequence – e.g. early up-bringing of children must antecede their educational performance. Similarly the family's social class may or *may not* influence a child's school performance, but a child's school performance can *never* influence a family's social class. In addition, as information accumulates from many investigations of a phenomenon, our confidence increases that the causal inference is correct. A notable topical example is that smoking is *probably* the cause of lung cancer.

'Hypothesis formulation', Part 1 of the block, starts with an introduction to the principles of research hypothesis formulation. It moves from an everyday commonsense exposition of the hypothesis concept via the scientific use of it to a brief examination of its logical structure. The use of research literature as a source of research hypotheses and research methodology is then considered and, finally, the transformation of research hypotheses into operational hypotheses which can be investigated empirically. Next the block presents some criteria by which the usefulness of research hypotheses can be judged, thus preparing you to examine some hypotheses yourself, and to work through an activity on hypothesis formulation in research to evaluate the ita method of teaching reading. The last section provides a brief summary of the main sections of a research design signposting the way to other sections of the course which treat some of the topics in greater depth. It then moves to the all important pilot investigation which should precede any properly conducted piece of educational research.

Part 2 of the block is concerned with sampling. The text examines the case for random sampling in survey research – explaining why the accuracy of estimates of the characteristics of a population is greatest when the sample of the population on which they are based is drawn by random sampling methods. The radio programme associated with Part 2 is 'Why Sample?' – a case study on the estimation of backwardness in reading in a London borough. Next, the text looks at the different types of sample design available to the researcher, ranging from the statistical ideal of the *simple random sample* to the *opportunity sample* employed in much educational research. An activity gives you practical experience of the different designs. The text then moves on to the crucial practical question of the optimum size of sample that the educational researcher needs to meet his research objectives and finally draws some general conclusions about the sampling problem in educational research.

The final section of the block introduces a topic which will be developed in greater detail in Block 4 – sampling statistics. The text looks at the sampling distribution with widest application in educational research – the *normal distribution* – and discusses the importance of two statistics which can be calculated for it – *the mean* and *standard deviation*. This leads to a definition of *sampling error* which underpins the whole subject of *statistical inference*. You work through some calculations of standard errors for educational research data in the text. Finally the importance of the sampling error concept in evaluating educational research is discussed. The block paves the way to the next crucial stage in the research process – methods of data collection – considered in Block 3.

Part 1
Hypothesis formulation

Notes on Reading and Working Order

READING

Set books

ENTWISTLE, N. J. and NISBET, J. (1972) *Educational Research in Action* (referred to as *ERIA*), London, University of London Press, Chapters 2 and 10.

OPPENHEIM, A. N. (1966) *Questionnaire Design and Attitude Measurement*, London, Heinemann, Chapter 2.

THORNDIKE, R. L. and HAGEN, E. (1969) *Measurement and Evaluation in Psychology and Education*, New York, John Wiley, Chapter 15.

WORKING ORDER

The main reading for Part 1 of the block is the correspondence text. Other reading is provided in the extract from Higson's article in Chapter 2 and Vernon's article in Chapter 10 in *ERIA*, Chapter 15 in Thorndike and Hagen and Chapter 2 in Oppenheim. You will probably find it best to read through the correspondence text quickly once and then work through it carefully, reading the other material as it is referred to, and finally doing the activity (p. 24) at the end.

Contents

		Page
	OBJECTIVES	10
1	INTRODUCTION	11
	The need for hypotheses	11
	Scientific reasoning	12
	The logic of scientific inference	12
	Formulation of hypotheses	14
2	REVIEW OF LITERATURE	15
	Sources	15
	Recording the material	16
	Choice of research instruments	17
3	FORMULATION OF HYPOTHESES	19
	Making the hypothesis operational	19
	Some criteria for hypothesis formulation	20
4	SUMMARY OF RESEARCH STRATEGY	22
	Four sections of a research design	22
5	PILOT WORK	23
	ACTIVITY	24
	ANSWERS	28
	REFERENCES	31
	ACKNOWLEDGEMENTS	31

Objectives

By the time you have completed this part of Block 2 you should be able to:

1. Define the term 'hypothesis'.
2. Say what place hypothesis formulation has in educational research.
3. Distinguish between hypotheses that are testable and those that are not testable.
4. Undertake a review of previous literature.
5. Select appropriate information from the review.
6. Organize this material.
7. Transform a general idea or problem into testable operational hypotheses.
8. Test the formulation of such hypotheses against given criteria.
9. Draw up a format of a research design.
10. Distinguish cases where a pilot study is necessary.

HYPOTHESIS FORMULATION

1 INTRODUCTION

The need for hypotheses

1.1 The last block described how research topics can arise and pointed out that the problem, still probably rather vague at this stage, has to be translated into precise operational hypotheses upon which a research plan can be designed. This planning stage is possibly the most demanding and certainly the most important part of the research process. Carrying out the research plan once formulated, i.e. collecting and analysing the data, is to a certain extent a mechanical process requiring, according to Van Dalen (1966) 'persistence rather than profundity'. But this does not mean that the researcher can step back into the wings once the research is on. He must be engaged continually in supervising the research process and adapting his procedures where necessary to take account of new situations and problems as they arise.

1.2 You might think that hypothesis formulation is necessary only in research of the experimental type. As we saw in Block 1, in such research the investigator proposes a hypothesis to explain some phenomenon and then designs an experiment to test it. In fact, hypothesis formulation is inherent in any research activity even of the descriptive kind. The particular descriptions we *choose* to make (questions in public opinion polls for example) are all made on the basis of suppositions we hold about the nature of the world. In the most general sense all such suppositions are hypotheses. A hypothesis is defined as *any* supposition or proposition assumed for the sake of argument. In John Stuart Mill's words 'A hypothesis being a mere supposition, there are no other limits to hypotheses than those of the human imagination.'

1.3 In science the word 'hypothesis' is generally used in a more restricted sense to refer to conjectures that can be used to *explain* observations. In practice these conjectures must satisfy certain conditions if they are to serve this purpose. An everyday example should serve to highlight some of these conditions. Suppose you went out to your garage one morning and found that the car would not start. You would no doubt ponder on the reasons for this. Try to think of three reasons and list them in the space provided.

1

2

3

1.4 There could of course be a large number of causes – a blockage in the petrol pipe, flat battery, condensation, dirt on the spark plugs and so on. Whichever you choose is an implied hypothesis, e.g. that dirt on the spark plugs

caused the failure to start. Your hypothesis can be tested by cleaning the plugs and attempting to start the engine again. If the engine does start you can be justified in saying that it is *probable* that dirty plugs *caused* the failure to start. If, on the other hand, the engine again failed to start, then you can say that the dirt on the plugs was not the cause, or at least not the *only* cause, of the starting problem.

Scientific reasoning

1.5 This is the usual sort of deductive reasoning we use when something unexpected happens, and two features of this process should be noted:

1 The hypothesis is not an unmotivated or undirected guess – it is designed to lead to an answer to a particular question (in our example, what caused the starting problem).

2 The hypothesis is such that, if true, it will have observable consequences – e.g. if dirt was the cause of the starting failure, then there will be dirt on the spark plugs which if removed will enable the engine to start. In other words the hypothesis can be *tested.*

1.6 However, simply because the predicted consequences of a hypothesis *are* observed does not *establish* that the hypothesis is correct. The test merely lends *support* to the hypothesis. It does not *prove* it. Let us return to our example. Although the plugs *were* dirty, the fault could have been caused by a faulty plug connection which was automatically remedied when the plugs were replaced. We cannot therefore claim categorically, on the basis of our observations, that dirty plugs caused the failure to start. The main point to remember therefore is that *hypotheses can be supported but not proved.* The term 'proved' carries the connotation of finality and absolute certainty. Hypotheses are never proved by producing factual evidence that is in harmony with their consequences; they are simply not disproved. On the other hand if the observed facts do not confirm the prediction made on the basis of a hypothesis, and the hypothesis has been tested adequately, then it is *conclusively rejected.* If cleaning the plugs *thoroughly* has no effect on the starting problem, then dirty plugs alone could definitely not have caused the problem. The hypothesis must have been wrong.

The logic of scientific inference

1.7 This mode of accounting for problems is the characteristic pattern of scientific thinking. It is, in fact, the hypothetico-deductive approach you read about in Block 1. It contains three essential steps:

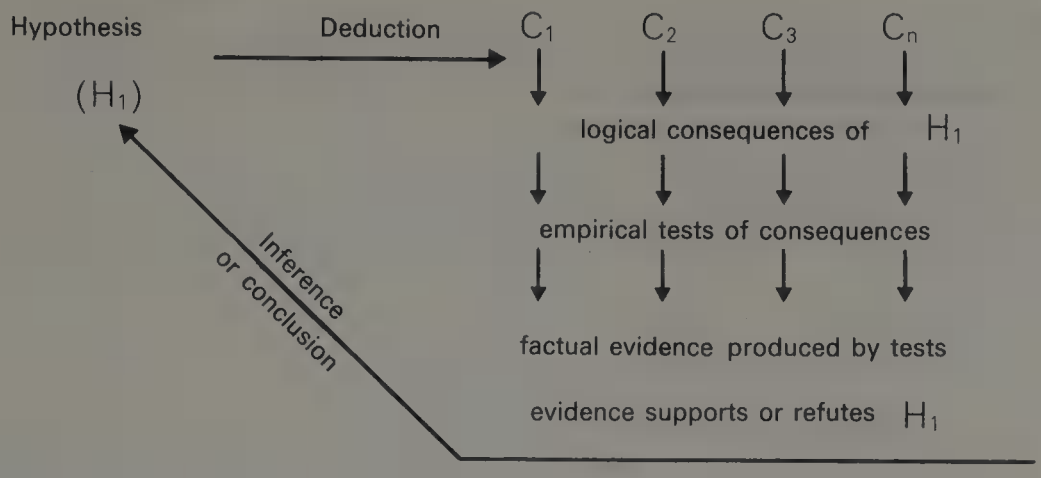
1 The proposal of a hypothesis to account for a phenomenon.

2 The deduction from the hypothesis that certain phenomena should be observed in given circumstances.

3 The checking of this deduction by observation.

1.8 Hypothetico-deductive arguments have a clear logical structure which can help one to see the pitfalls that the unwary researcher may fall into. We start with a supposition or assertion H_1 which has a number of deducible consequences – C_1, C_2, C_3. By testing these consequences by appropriate empirical procedures (e.g. educational measurements) a conclusion is drawn. This is used to confirm or refute the hypothesis. The diagram on p. 13 (Fig. 1) taken from Van Dalen illustrates this process.

Fig. 1 The process of testing a hypothesis



1.9 But notice the logical status of the conclusion drawn. It is an *inference*, not a *proof*. The researcher who tests and confirms one consequence of a hypothesis may think that he has established the cause of the phenomenon he is investigating but in doing this he could be making a basic logical mistake. His argument can be analysed logically as follows:

Major premise	If H_1 (antecedent) is true then C_1, C_2, C_3 . . . C_n (consequent) follow.
Minor premise	C_1, C_2, C_3 are true (empirical evidence affirms the consequent).
Conclusion	Therefore H_1 is true (inferred fact).

However, this reasoning is faulty; formal logic insists that if a minor premise *affirms* the consequent, as it does in this instance, then the hypothetical argument is *not* valid. The researcher may show that C_1, C_2, C_3 support H_1 or confirm it to a certain extent, but *other* hypotheses may also account for these consequences. What the evidence C_1, C_2, C_3 can do is *refute* the hypothesis. If one single consequence of it turns out to be unsupported then the hypothesis must be wrong. This distinguishes the scientific hypothesis from everyday speculation and conjecture. A *scientific* hypothesis is one that can be disproved. We shall see what this implies later.

1.10 Let us look at an educational example. An educational researcher may have reasoned that poor family background causes poor reading attainment in children. He may have tried to produce empirical evidence that low family income and overcrowding are associated with poor reading attainment. If no such evidence was forthcoming then, as we have seen, the hypothesis must be decisively rejected. But if the predicted relationship was found could the researcher conclude that his hypothesis was correct – i.e. that poor family background does cause low reading attainment? The answer must be 'no'. It might equally have been the case that bad teaching is to blame which may also be associated with the consequences of a poor home background, because schools in poor neighbourhoods may attract bad teachers. In other words there may be alternative hypotheses which explain the phenomenon of reading backwardness equally well. These alternative hypotheses will have other deducible consequences, e.g. if poor teaching in poor neighbourhoods causes backwardness in reading, then inservice training of teachers in schools in poor neighbourhoods will improve reading attainment among the children in these schools. To underline the main point again, the scientific process never leads to *certainty* in explanation – only the rejection of existing hypotheses and the construction of new ones, which stand up best to the test of empirical evidence.

Formulation of hypotheses

1.11 But how do we formulate hypotheses in the first place? One of the distinctive features of scientific enquiry is that the researcher does not generally formulate his hypotheses until he is thoroughly familiar with the theoretical and empirical background to the problem. This does not mean, however, that the hypotheses always emerge clearly from this body of empirical knowledge once the researcher is familiar with it. On the contrary, the ideas which stimulate an educational enquiry may emerge in a rather haphazard albeit often creative and imaginative way, as the following quotation from Gage (1963) suggests.

> Actually, the process of doing research . . . is a rather informal, often illogical and sometimes messy-looking affair. It includes a great deal of floundering around in the empirical world. . . . Somewhere and somehow in the process of floundering, the research worker will get an idea. In fact, he will get many ideas. On largely intuitive grounds he will reject most of his ideas and will accept others as the basis for extended work. . . . If the idea happens to be a good one, he may make a significant positive contribution to his science – 'may' because between the idea and the contribution lies a lot of persistence, originality, intuition, and hard work. It is in this sort of activity, rather far removed from the public, more orderly and systematic phases of scientific work, the productive research worker spends much of his time and effort. . . . The process of getting and developing ideas is undoubtedly a confused mixture of observation, thinking, asking why, cherishing little unformed notions, etc. . . . But it should also be emphasized that the notion cannot be allowed to remain forever vague: ideas must eventually achieve clarity and testability if they are to receive serious attention from other researchers.

1.12 Ideas achieve 'testability' when they are transformed into testable hypotheses. But in practice researchers do not contemplate the outcomes of their studies without a thorough understanding of the basic body of facts in the relevant area. This understanding identifies the major theoretical concepts and controversies. Confirmed theories are, in the strictest sense of the word, logically connected sets of established hypotheses, whose major functions are to describe, predict and explain their subject matter. As such, a good theory can be regarded as both a summary of past experience and a predictor of future experience. Research which is not based on a review of relevant literature thus may become an isolated study which is at worst fundamentally unsound and at best bears only accidental relevance to what has gone before.

1.13 This is assuming of course that there is previous research to go on; in some new problem areas this may not always be the case. A phenomenon like drug-taking in school children, for example, may demand new insights which earlier research cannot provide. Moreover, the formulation of hypotheses certainly is not a mechanical process which relies entirely on past theory and research. Past theory and research are important, but must be balanced with the imaginative insights of the investigator, as the following passage from Van Dalen (1966) indicates:

> A scientist cannot be an isolated rebel who relies on personal observation alone to create fruitful hypotheses. His work will be hopelessly hampered if he is not thoroughly familiar with established facts, existing theories, and previous research relating to his problems. While manipulating relevant raw materials drawn from these sources, he may locate the key associations or missing data needed to explain puzzling phenomena. Supplying the imaginative concepts – missing pieces – needed to solve a problem is his ultimate objective, but these concepts may be produced more readily if he is familiar with the known facts and orders them imaginatively. Familiarity with the literature in the field relating to a problem gives the trained person a tremendous advantage over the neophyte or outsider in arriving at a successful solution. Complete slavery

to traditional thought and excessive specialization in a field, however, may crush creativity.

1.14 The best path for the researcher to take, therefore, is the middle way between getting bogged down in previous work to the exclusion of all else, and wandering off along any track that his free ranging imagination or armchair speculation takes him to. Ignorance, however, is never a good position to start from. An essential for any researcher is to familiarize himself thoroughly with the existing body of research and theory in his field of study. To do this he needs to 'review the literature'. We shall consider this stage of the research design next.

2 REVIEW OF LITERATURE

Sources

2.1 The review of literature is normally undertaken in two stages. The first stage involves a general overview of the relevant area. This is followed by a more specific and structured review involving the salient research studies. Secondary sources are consulted in the initial review, i.e. general textbooks which include the relevant topics. Examples were mentioned in the previous block in the context of nursery education. In addition a number of books have recently been published with the expressed purpose of providing scholarly overviews of contemporary problem areas in education. Among these is a series entitled *Educational Research in Britain* edited by Butcher and Pont (1968, 1970, 1973). Three volumes have now been published, all of which contain chapters written by experts in their respective fields. These provide a synthesis of the major studies, outline the problem areas and suggest avenues for further research. A similar publication is *A Map of Educational Research* by Thouless (1969). Although these publications are limited to research in Britain, comparable texts exist which review American research. Among the most important are *The Encyclopedia of Educational Research* and *Review of Educational Research* (both published by The American Educational Research Association), the former being more comprehensive. Large numbers of such texts and research journals exist for all manner of educational enquiry. The article in Chapter 2 in *ERIA* by Higson (1968) provides a comprehensive review of these literature sources. You might find it useful to read it now.

2.2 Other sources are also useful. One text recommends that the student doing a research study should attach himself to a professor who is already engaged in research. Not much likelihood of that for an Open University student! However, the text does emphasize the value of talking or otherwise communicating with others engaged on similar projects. Conferences serve a useful function in this respect because they provide a meeting place where interested people can talk about matters of common concern. Most professional bodies have annual and regional conferences. The British Psychological Society, for example, has an educational section conference each year at which research papers are read. Such papers are usually more up to date than published studies.

2.3 The initial review indicates the central issues involved and serves to focus attention onto a more specific subtopic of the initial problem, and the hypotheses which have been advanced in relation to it. Suppose, for example, you were concerned with the general problem of the effect of personality on school attainment. The initial review would have shown that a number of theories of personality exist, that a number of relevant measures of personality and

attainment are available, that the problem has been attacked from a number of different perspectives and that a number of hypotheses are in vogue. It will have shown that relationships between measures appear to vary depending on age, sex, type of schooling and so on of the children who were studied. (You find a review of this kind in Unit 5 of *Personality Growth and Learning*, E281.) The sort of hypothesis which is currently holding its own is that the personality factor extraversion aids the school attainment of girls in the first years of secondary school but hinders the attainment of boys. Similarly in research on reading, the review will have shown that, among other factors, past research has considered the effect of intelligence, sex, impaired hearing, dyslexia, lack of sensory experience and the efficacy of different reading schemes. On the subject of attitudes and learning, the theme of the project for the course, relationships again have been shown to vary with age and sex. You will find some relevant research material on this topic included with this block.

Recording the material

2.4 Once this initial review has been completed it is useful to write a summary of the material sampled, outlining the major theories, salient studies and their results. Such a synthesis not only helps to clarify the major issues in the mind of the researcher but provides a foundation for further reading. At this stage a decision is made, if one has not been made already, about which particular aspect of the wider problem will be studied further. To take the personality example again, you may have become intrigued by the possibility that different school environments could affect the manner in which personality relates to school attainment. Having decided that this is the specific area of study which interests you, you embark upon the second stage of the review; this time you concentrate on primary source materials, i.e. research reports written at first hand. Normally there is little difficulty in knowing which reports are relevant and available since the major studies will have been indicated in the initial review. Each book or research article usually begins with a summary of past research and prints the references in full at the end.

2.5 The great amount of material gathered during the review requires some form of organization. The researcher usually records this information on a bibliography card. This card carries the author's name, date of publication, the title of the book or name of journal; it also contains a brief summary of the salient details of the research including the page numbers, title of the article, details of the sample, techniques of measurement and summary of results. The author's main conclusions and any weaknesses in them will also be recorded together with additional references gained from the article. And of course if the research is one of the major ones in the field of study, further notes on additional cards or in a notebook will also be taken. An example of a completed bibliography card is shown in Fig. 2.

2.6 One of the main aims of a research review is to evaluate critically the different researches in order to identify those findings on which most weight can be put, and to see what hypotheses might be put forward to explain apparently contradictory findings. As differences in research procedures can radically affect the results obtained from a research, accurate details of samples and measurement techniques are required. Details of the sample population in terms of size, age, sex, intellectual level, social class level, country of origin, and even whether it consists of urban or rural dwellers can all substantially affect the outcome. Details of measurement are very important because different measures are often used to assess the same attribute, often with conflicting results. In the Haddon and Lytton research used in the example of a bibliography card (Fig. 2) the measures used were derived from the Minnesota Tests of Creative

Fig. 2 Example of completed bibliography card

HADDON, F.A. and LYTTON, H. (1968) BRITISH JOURNAL OF EDUCATIONAL PSYCHOLOGY, 38, 171-180

'Teaching approach and the development of divergent thinking abilities in primary schools'

SAMPLE: 211 11-12 year olds from 4 schools, two designated 'informal' and two designated 'formal' by college of eduaction lectures. Schools matched for socio-economic background. Mean Verbal Reasoning quotient 101.

MEASURES: Verbal Reasoning - Moray House. 3 verbal and 3 non verbal divergent thinking tests adapted from MTCT. Sociometric test (given in appendix).

RESULTS: Children in informal schools significantly superior in divergent thinking abilities. Divergent children more popular in informal schools.

COMMENTS: Sample size of schools very small and unrepresentative.
Terms 'informal' and 'formal' open to different interpretation.

REFS: KELLMER PRINGLE, M.L. and MCKENZIE, I.R. (1965) Brit. J. Educ. Psychol. 35, 50-59.
Teaching method and rigidity in problem solving.

Thinking (MTCT)[1]; a number of other tests also claim to measure creativity/divergent thinking, and yet the correlation (strength of relationship) between these tests can be low. Thus it may be found that studies which have set out to investigate the same hypothesis, but with different measures, obtain different results. Also the same attribute can be measured by different techniques. In measuring personality traits, for example, use can be made of self-rating questionnaires, observer ratings, objective tests or projective tests, e.g. interpretation of inkblots, etc. You will find examples of the latter type in Thorndike and Hagen, p. 494. Accurate details of all the different measures employed in the research area will be of great help when the researcher makes his decision as to which research instrument(s) to use himself.

Choice of research instruments

2.7 A number of factors have to be taken into account when deciding which alternative measure or measures are to be chosen for a study. We shall be looking in detail at these factors in Block 3 but a preliminary consideration is in place here. Remember that the quality of data collected relies on the *quality* of the measures adopted. The two principal features of this quality are *reliability* and *validity*. Reliability concerns the consistency or repeatability of a measure. A perfectly reliable measure is one which if administered twice under the same conditions would provide identical results. Thus, if a test of intelligence or personality was administered twice to the same group within a short space of time, and the results of the two testings varied greatly, then the reliability of the test would be suspect. Validity, on the other hand, refers to the extent to which the instrument measures what it purports to measure. For example, to what extent do scores on tests of creativity indicate creative performance in real life situations?

[1]An example of an item from this test is: How many unusual uses of a shoebox can you think of?

2.8 The time and expertise required in the administration and scoring of measuring instruments are also important factors in the decision-making process. A personality questionnaire can be administered to groups of subjects in a short time, and can be quickly scored with a template by inexperienced clerical workers. A projective test of personality, on the other hand, often requires individual testing and must be scored by trained personnel; the time and expertise required are therefore much greater.

2.9 The possible biases introduced into research by different administration and scoring procedures for the tests used are highlighted in the study of 'divergent thinking' (creativity). As Vernon (1971) has reported in an article (part of which is included in Chapter 10 of *ERIA*) virtually every investigator in this field has employed different scoring methods for the tests employed. In addition there has been a marked difference of opinion about the conditions under which these tests should be administered.

2.10 Many investigators administer these tests formally, i.e. to groups, under test conditions, with time limits imposed. Others argue that such tests should be administered informally, i.e. individually, in a relaxed or gamelike atmosphere, and without time limits. A typical test item of this type would be 'list as many uses as you can think of for a brick (or barrel, or piece of string etc.)'.

SELF-ASSESSED QUESTION 1

Indicate in the space provided:

1 Which procedure – formal or informal – do you think would elicit the greater number of responses?

---

2 Which procedure would take a greater amount of time?

---

3 What other problems might occur in the informal situation?

You will find it useful to compare your answer with what Vernon reports in the extract from his article reproduced in Chapter 10 of *ERIA*.

2.11 At least five factors should therefore be taken into account when reviewing the measuring instruments used in earlier studies. They suggest questions which the investigator needs to answer when deciding which instruments to use to test his own hypotheses.

1 *Reliability* How reliable is the measuring instrument and how does its reliability compare with other possible instruments for measuring the same attribute?

2 *Validity* What evidence is there that the measuring instrument does provide a good assessment of the attribute, as defined by the investigator?

3 *Time to administer* How long does the measuring instrument take to administer and is there sufficient time to use it within the confines of the study?

4 *Expertise required* Does the investigator require special expertise in order to use the instrument or can it be used with standard instructions by anybody?

5 *Administration procedures* What biases are likely to be introduced into the subjects' responses depending on the administration procedures employed and the situation in which the instrument is used?

3 FORMULATION OF HYPOTHESES

3.1 Having completed the review stage of the research design we should now be in a position to put the general problem or idea into a form which enables us to investigate it. To take the example concerning personality again, we may have decided to investigate Eysenck's theory of personality in relation to school attainment in eleven-year-old children – i.e. to what extent can Eysenck's theory help to explain good or poor attainment in these children? The research hypothesis will then be made more specific. It might be expressed as *'extraversion' and 'neuroticism'*[1] *are related to school attainment in eleven-year-old children.* This hypothesis is still not in a form which can be tested. What method of measuring personality is to be used? To what extent are neuroticism, extraversion and attainment related? The research hypothesis therefore needs to be operationalized – i.e. the phenomena referred to in it need to be *defined* in terms of observations that can be made. (You will remember that we looked at the concept of operationalization briefly in Block 1.)

Making the hypothesis operational

3.2 To make a hypothesis operational we need to *define* the phenomena to which it refers in terms of *observations*. In practice this means expressing the hypothesis in terms of *variables* that can be *measured* – a variable being any entity (characteristic, attribute, phenomenon) that can take on different values. One of the main problems in social science generally, and educational research in particular, is that many variables we wish to investigate are very difficult to operationalize – we have no good means of measuring them. For example even an apparently simple variable like attainment may be very difficult to assess, if it refers to a developing skill such as critical understanding or creative writing. What we are forced to do is make the compromise between the ideal measure of the variable we wish to study and available measures of it, such as attainment tests and teacher ratings. In other words we operationalize the variable in terms of the measures we have available or in terms of the measures we develop specifically for the purpose.

[1]Eysenck's two principal dimensions of personality.

3·3 To return to our research hypothesis about the relationship between personality and school attainment, suppose that the measures used were the Junior Eysenck Personality Inventory (the JEPI), which is designed to measure the variables 'extraversion' and 'neuroticism' and a standardized test[1] of English which is designed to measure the variable English attainment. Write in the space below two operational hypotheses which could be derived from the research hypothesis.

3·4 In this instance you could have postulated that 'Extraversion, as measured by the Junior Eysenck Personality Inventory, will be related *positively* to English attainment, as measured by standardized test X', i.e. the higher a child's JEPI extraversion the *higher* will his score be on test X. *Or* 'Neuroticism as measured by the Junior Eysenck Personality Inventory will relate *negatively* with English attainment measured by standardized test X', i.e. the higher a child's JEPI neuroticism score the *lower* will his score be on text X. In this case the expected direction (negative or positive) and strength of relationship will have been indicated by the review of the literature, although this is not always the case.

Some criteria for hypothesis formulation

3·5 It is essential that hypotheses be carefully formulated. Travers (1969) suggested five criteria to which they should conform:

1 *Hypotheses should be stated clearly and in correct terminology.*
General terms such as personality, school attainment, social class etc. should be avoided. The statement demands concise technical language and definition of terms. More appropriate therefore are – 'personality as measured by the Junior Eysenck Personality Inventory'; 'school attainment as measured by English Progress Test E (National Foundation for Educational Research)'; 'social class as defined by the Registrar General's classification of head of household's occupation'[2] etc.

2 *Hypotheses should be testable.*
Since hypotheses are predictors of the outcome of the study, an obvious necessity is that instruments should exist (or can be developed) which will provide valid measures of the variables involved.

3 *Hypotheses should state relationships between variables.*
A satisfactory hypothesis is one in which the expected relationship between the variables is made explicit. If you refer back to the personality and

[1]The term 'standardized test' refers to a test for which 'national norms' are available. These are the scores which were gained by the sample of the population for which the test was standardized. For example the tests which used to be used for selection at eleven plus are of this type. The standardization data for the test states the scores which one can expect children of different ages to obtain. The tests teachers develop for their end of term tests are of a different type since they are normally only relevant to the individual class.

[2]UK General Register Office (1970) *Classification of Occupation*, London, HMSO.

attainment example, you will note that the variables involved are extraversion or neuroticism as measured by the Junior Eysenck Personality Inventory and attainment as measured by a standardized test of English. The relationship posited is that the personality variables and the attainment variable will be related positively or negatively.

4 *Hypotheses should be limited in scope.*

Hypotheses of global significance are not required; those that are specific and relatively simple to test are preferable. It is of course possible to state a rather broad research hypothesis and derive a number of operational hypotheses from it.

5 *Hypotheses should not be inconsistent with most known facts.*

All hypotheses should be grounded in past knowledge, i.e. the knowledge gained from a review of the literature. They obviously cannot be consistent with *all* known facts since many studies give contradictory results. And in these cases the hypothesis may be formulated to resolve the contradiction. For example, differences in the relationships found between personality and attainment in different age groups might be due to methodological errors or might be explained by a new hypothesis that the *form* of the relationship changes with age. There is also certainly room for the use of imagination in the extrapolation of known facts to new populations.

SELF-ASSESSED QUESTION 2

Now discuss the following examples in terms of the five criteria above. Look at each hypothesis and say whether it satisfies all the criteria; if it does not, say why not.

Example 1

Among ten-year-old junior school children, extraverts, as measured by the Junior Eysenck Personality Inventory, will gain significantly higher scores on a standardized test of English than will introverts.

Example 2

Progressive teaching methods have led to a decline in academic standards in primary schools.

Example 3

The introduction of social sciences into the curricula of secondary schools will produce better citizens.

Example 4

The scores gained on a test of academic motivation by ten-year-old children from Social Class I (as defined by the Classification of Occupations 1970) will not be significantly different from those gained by children from Social Class V.

Compare your answers with those on pp. 28–9.

4 SUMMARY OF RESEARCH STRATEGY

4.1 After the hypotheses have been formulated a research design must be constructed so that they can be tested. The design is an inventory of what must be done and what materials and techniques are necessary for the implementation of the study. In the report of the research this plan is described in detail under the general heading of *research design*. A research design usually has four sections, each involving a different aspect of the study. Only a brief description of these sections is given here because we have already considered the first and the others will be dealt with in greater detail in later blocks.

Four sections of a research design

Section 1 – General introduction

4.2 In the introduction the problem is summarized and a concise account of the theoretical background is included. The hypotheses are stated, as are the assumptions underlying the study. These assumptions are important since many attributes are measured indirectly. If inkblots are to be used as a measure of personality, as in the Rorschach test (Thorndike and Hagen, p. 495) then a rationale for the acceptance of such an assumption is necessary. Another example is contained in the Haddon and Lytton (1968) study already referred to. Many educational researchers make the claim that individuals who score

highly on tests of divergent thinking are 'creative'. Haddon and Lytton discuss this assumption in their introduction:

> Despite the attention which creativity has received as a fashionable topic for research there is as yet no clear evidence for the assumption that high scores on tests of divergent thinking will be particularly fertile in creative, original production in their own life situation . . . . To avoid undue claims we have used the term 'divergent thinking' rather than 'creativity' throughout.

Similarly, if the study concerns backwardness in reading, for example, it must be made clear by what criteria the backward readers are to be defined. In other words the terms must be operationalized; e.g. the backward reader is one who fails to gain a certain score on a specified reading test.

Section 2 – Statement of measures

4.3 The review of literature will probably have delineated the tests and other measures suitable for the age level under consideration. There are advantages in using measuring instruments which have previously been successful for the reasons discussed earlier. (We shall be looking at the whole subject in detail in Block 3.) In many studies evidence for reliability and validity of the measuring instruments are given which may give the researcher confidence in them. In addition, the use of established test instruments spares the researcher the arduous and time-consuming necessity of developing and validating new instruments. Finally, continuity between past and present research is established. The new study can contribute to an increasing body of knowledge in a meaningful way. If the variables to be investigated have been assessed previously by standardized tests, their use also carries advantages. Scores of the study sample can be compared with those obtained nationally. This comparison may also give some indication of whether the sample investigated is representative on that particular measure.

Section 3 – The population to be studied

4.4 The population has probably already been defined by the hypotheses; thus in the personality example the population was eleven-year-old schoolchildren. Obviously we cannot study all eleven-year-old schoolchildren; instead we select a portion of this population for study. This portion is called the sample. Ideally the sample is chosen so that it is representative of the population, i.e. is a *perfect miniature* of it. This is achieved by what is called 'random sampling'. Random sampling, and the type and size of sample, will be discussed in detail in the next part of the block. Briefly, the type and size of sample chosen depend on the type of study, the amount of finance, time and assistance available and, not least, on the availability of subjects. It is one thing to decide on sample size and another to achieve it. The size of sample therefore varies from a national survey, involving thousands of schoolchildren, to a case study of a single child.

Section 4 – Methods of processing data

4.5 It is essential that the researcher has a clear idea of how the data are to be analysed *before* they are collected. A check should therefore be made that the type of data to be collected can be analysed in the way proposed, e.g. by computer. We shall be looking at methods of data analysis in Block 4.

5 PILOT WORK

5.1 Creating a research design is often a complex business. Certain measuring instruments may not exist, or existing ones may have poor reliability or

inadequate evidence of validity. A number of measures may exist which supposedly assess the same attribute, but which have not been tried at the age level required. Many studies use questionnaires or attitude scales which have been specially developed for that study, and these have to be assessed to find out if they are suitable. It is therefore standard practice to carry out a *pilot* study or series of pilot studies to assess the effectiveness of all measures before the actual study. A pilot study is a small-scale replica of the main study, a dress rehearsal, carried out ideally on a random sample of the study population. It may follow extensive exploratory work with the members of the age group under consideration before the measuring instruments are developed and tested on any large scale. Chapter 2 in your set book by Oppenheim (1966) contains a useful account of the procedures involved.

5.2 The principal function of a pilot study is to assess the adequacy of the measuring instruments and administration procedures. A relevant example of this concerns the Open University project on 'attitudes to school'. The course team found a number of scales which purported to measure a general attitude to school variable or 'academic motivation' but a number of these related only to primary schools and others to secondary schools. The relationships between the different measures were not known, and there was insufficient evidence of reliability and validity for some scales. An initial pilot study was therefore essential so that these points could be checked, and any amendments to the scales made. They were then piloted a second time to try out administrative procedures for students to use in the main study.

5.3 Pilot studies are also valuable in that the efficacy of the hypotheses can be assessed, particularly those derived from tenuous data, and questions about sample design, such as the optimum sample size, can be resolved. Finally, if a random sample has been used, the researcher should ask himself whether, in the light of the results gained, the main study is worth carrying out as planned. Adequate pilot work is therefore the last safeguard against the possibility that the main study may be ineffective.

ACTIVITY

Hypotheses on the effect of ita

This activity is designed to take you through the process of formulating hypotheses, and you should spend about thirty minutes on it. We have selected a well-documented project for you to work through, and you can first generate your own hypotheses and then compare them with those actually researched. The project is the investigation of learning to read by means of the initial teaching alphabet (ita), which was also the subject of your first radio programme. Sir James Pitman designed ita not only to reduce the many anomalies of English spelling, but also to ensure that transition to traditional orthography (to) should be as easy as possible once some level of reading fluency was achieved. Here is a sample paragraph of ita:

> ðis is printed in an augmented rœman alfabet, ðe purpos ov whiçh it not, it miet bee suppœsd, tœ reform our spelliŋ, but tœ imprœv ðe lerniŋ ov reediŋ. it is intended ðat when ðe beginner has açheevd ðe iniʃhal suksess ov flœensy in ðis speʃhally eesy form, his fuetuer progress ʃhœd bee konfiend tœ reediŋ in ðe present alfabets and spelliŋs ov ðem œnly.
>
> if yœ hav red as far as ðis, ðe meedium will hav prœvd tœ yœ several points, ðe mœst important ov whiçh is ðat yœ, at eny ræt, hav eesily mæd ðe çhænj from ðe ordinary rœman alfabet with konvenʃhonal spelliŋs tœ augmented rœmen with systematik spelliŋ.

In October 1960 John Downing was appointed Director of the Reading Research Unit, University of London, with the task of testing whether children could learn to read more effectively with an augmented alphabet. From the research viewpoint this meant testing whether the irregularity of application of the English alphabet was a serious handicap in learning to read, in comparison with a contrasting medium – the initial teaching alphabet. Now before reading any further take about five minutes to write down (in the space provided) a few *questions* that you would want to ask in order to assess the effectiveness of any augmented alphabet – or i t a in particular – in teaching children to read.

At the beginning of Downing's report on the i t a experiment he summarizes the questions to which his team addressed itself. These were:

1 Can children learn to read more easily with i t a than they can with t o?

2 Can pupils transfer their training in reading i t a to reading in t o?

3 After the whole process of beginning with i t a and transferring to t o, are reading attainments in t o superior to what they would have been without the intervention of i t a?

4 Will children's written compositions be more fluent with the simpler i t a code for speech? (That is, will the gap between their spoken and written vocabularies be narrowed?)

5 How will children's later attainments in t o spelling be influenced by their earlier experiences of reading and writing the different spellings of i t a?

Refer back to your own questions and compare them to Downing's.

Are there any questions which you feel to be important which Downing has not investigated? Are there practical reasons why he might not have investigated these issues? (There might be financial, ethical or time limitations; the issues might be of secondary importance, or outside Downing's brief – e.g. he was not asked to investigate whether Pitman's ita was the best possible form of transitional alphabet.)

Now write down in the space below one or more operational hypotheses for each of Downing's questions, and then check that each one meets the five criteria laid down in section 3. You should spend about ten minutes on this.

Question	Hypothesis
1	
2	
3	
4	
5	

Now read the opening of Downing's report (pp. 27–8) and see what hypotheses he himself tested in relation to each question.

You will notice that these are not all stated in operational terms at this point in his report. Indicate by each of them whether they are stated operationally or not by ticking the appropriate boxes on p. 27.

Hypothesis	Not operational	Operational	Notes
1			
2			
3			
4			
5			
6			
7			

Compare your answers with those on p. 30.

Chapter 1 '*The ita symposium*' (NFER, 1967)

NOTE: Please do not read this excerpt until the activity tells you to do so.

Aims of the first ita experiment

In September 1960, a design for an investigation of the effectiveness of the Initial Teaching Alphabet for beginning reading was prepared. Because the scope of the evaluation programme of ita had, necessarily, to be limited by the funds available, it was decided to confine the investigation to the study of a few basic, practical questions which appeared to be of greatest importance to the teachers and administrators who were considering using ita in infant schools. A full statement of these questions and the related hypotheses will be published elsewhere[1] but for the present purposes they may be summarized as follows:

[1]DOWNING, J. A. (1967) *Evaluating the Initial Teaching Alphabet*, London, Cassell.

Question 1: Can children learn to read more easily with ita than they can with traditional orthography? Two major hypotheses are related to this question:

Hypothesis 1: Because of the reduction in the volume of learning in ita as compared with to, children in ita classes should make significantly more rapid progress through their basic reader series.

Hypothesis 2: Because of the greater regularity of grapheme-phoneme relations in ita as compared with to, pupils learning to read with ita should achieve significantly higher scores on reading tests in which lower-order decoding skills have an especially important role to play.

Question 2: Can pupils transfer their training in reading ita to reading in to?

Hypothesis 3: Because of ita's design as a transitional alphabet, in ita classes, reading achievements in to should not be inferior to previous achievements in ita once fluency in ita has been established.

Question 3: After the whole process of beginning with ita and transferring to to, are reading attainments in to superior to what they would have been without the intervention of ita? The major hypothesis here is that:

Hypothesis 4: Through positive transfer from ita learning to to learning, pupils who have first learned to read with ita and then made the transition to to should read the latter with significantly greater accuracy, speed and comprehension than pupils who have not used ita in the beginning.

Question 4: Will children's written compositions be more fluent with the simpler ita code for speech? (That is, will the gap between their spoken and written vocabularies be narrowed?). Two hypotheses were to be tested here:

Hypothesis 5: Because ita also simplifies and regularizes the encoding operation, the written compositions of ita pupils should be longer than those of children who begin to read in to.

Hypothesis 6: The written vocabulary of ita pupils should be more extensive than that of their to counterparts.

Question 5: How will children's later attainments in to spelling be influenced by their earlier experiences of reading and writing the different spellings of ita? The early experiences with the regular grapheme-phoneme relations of ita should also produce positive transfer to to spelling. Therefore, the hypothesis is that:

Hypothesis 7: Spelling attainments in to after the transition stage should be superior in classes where ita was used for the beginning stage.

ANSWERS

Self-assessed question 2, pp. 21–2.

Example 1

Among ten-year-old junior school children, extraverts, as measured by the Junior Eysenck Personality Inventory, will gain significantly higher scores on a standardized test of English than will introverts.

In this example the hypothesis is formulated in concise language in which the terms are defined: there is a stated relationship between the variables of

extraversion, introversion and English attainment, which is testable; it is not too broad in scope, and is consistent with most known facts. All five criteria are therefore satisfied.

Example 2

Progressive teaching methods have led to a decline in academic standards in primary schools.

Two variables – progressive teaching methods and academic standards – are related, thus satisfying the third criterion. However, general terms are used – how are progressive teaching methods to be defined? Can they be tested? What academic standards are involved? Over what period of time is the comparison to be made? Are all age groups within primary schools involved? What evidence is there for this decline? It is clear that only criterion 3 is satisfied in this example.

Example 3

The introduction of social sciences into the curricula of secondary schools will produce better citizens.

The major disadvantage of this hypothesis, in its present form, is that it is virtually untestable. To find out whether social science courses and citizenship are related would require that pupils who have taken social science courses have grown to adulthood, criteria of citizenship would have to be formulated against which each subject could be evaluated and, most difficult, it would have to be determined what part the social science course had played in each individual's citizenship rating. Furthermore, terms such as 'better citizen' involve value judgements which are unlikely to be acceptable to all investigators – i.e. any one operational definition of the term is unlikely to satisfy many people.

Example 4

The scores gained on a test of academic motivation by ten-year-old children from Social Class I (as defined by the Classification of Occupations 1970) will not be significantly different from those gained by children from Social Class V.

There would appear to be no criterion which this hypothesis does not satisfy. In addition by being stated in what is called the null form it is amenable to statistical testing – the main topic of Block 4.

Answers to activity, p. 27.

Downing's hypotheses

None of Downing's hypotheses are operational in the strictest sense, because he does not set precise limits on any of the measurements he proposes. But some of them come much closer to the operational form than others as indicated on p. 30.

Hypothesis	Not operational	Operational	Notes
1	✓		'Rapid progress' not operationally defined.
2		✓	Type of test required is operationally defined.
3	✓		'Reading achievements' and 'fluency' not operationally defined.
4	✓		Comprehension' not operationally defined.
5		✓	Length of essay is an operational measure, i.e. number of words.
6		✓	Extremeness of vocabulary is number of difficult words the child uses.
7		✓	Number of spelling mistakes is an operational measure.

References

BUTCHER, H. J. and PONT, H. B. (1968, 1970, 1973) *Educational Research in Britain*, London, University of London Press.

GAGE, N. L. (1963) *Handbook of Research on Teaching*, Chicago, Rand McNally.

HADDON, F. A. and LYTTON, H. (1968) 'Teaching approach and the development of divergent thinking abilities in primary schools', *Brit. J. Educ. Psychol.*, 38, pp. 171–80.

OPPENHEIM, A. N. (1966) *Questionnaire Design and Attitude Measurement*, London, Heinemann.

NISBET, J. D. and ENTWISTLE, N. J. (1972) *Educational Research in Action*, London, University of London Press.

THORNDIKE, R. L. and HAGEN, E. (1969) *Measurement and Evaluation in Psychology and Education*, New York, John Wiley.

THOULESS, R. H. (1969) *A Map of Educational Research*, Slough, NFER.

TRAVERS, R. M. W. (1969) *An Introduction to Educational Research*, New York, Macmillan.

VAN DALEN, D. B. (1966) *Understanding Educational Research*, New York, McGraw-Hill.

VERNON, P. E. (1971) 'Effects of administration and scoring on divergent thinking tests', *Brit. J. Educ. Psychol.*, 41, pp. 245–57.

Other sources

BORG, W. R. (1963) *Educational Research: An Introduction*, New York, David McKay.

O'CONNOR, D. J. (1957) *An Introduction to the Philosophy of Education*, London, Routledge and Kegan Paul.

FOX, D. J. (1969) *The Research Process in Education*, New York, Holt, Rinehart and Winston.

Acknowledgements

Grateful acknowledgement is made to the following for material used in Part 1 of this block:

McGraw-Hill Book Company for D. B. Van Dalen, *Understanding Educational Research*, Copyright © 1966 McGraw-Hill Book Company and used with their permission; NFER Publishing Company Ltd. for J. Downing, *The ita Symposium.*

Part 2
Sampling

Notes on Reading and Working Order

READING

Set Reading

NISBET, J. D. and ENTWISTLE, N. J. (1972) *Educational Research in Action* (referred to as *ERIA*), University of London Press, Chapter 3.

THORNDIKE, R. L. and HAGEN, E. (1969) *Measurement and Evaluation in Psychology and Education*, New York, Wiley.

Recommended reading

GUILFORD, J. P. (1965) *Fundamental Statistics in Psychology and Education*, New York, McGraw-Hill, Chapter 4.

MOSER, C. A. and KALTON, G. (1971) *Survey methods in social investigation*, London, Heinemann.

WORKING ORDER

Section 2 of the correspondence text should be read in conjunction with Chapter 3 in *ERIA*.

Section 3 of the text requires the student to read parts of, and do computation from, Chapter 5 in Thorndike and Hagen (third edition). Students already with some statistical expertise should also read Chapter 4 in Guilford. Those who want to pursue the subject of sampling design in greater depth might like to read the excellent account by Moser and Kalton (1971).

Graph paper is supplied with the block.

Contents

		Page
	OBJECTIVES	36
1	INTRODUCTION	37
	The need for sampling	37
	The random sample	38
2	SAMPLING DESIGNS	41
	Systematic sampling	41
	Stratified sampling	42
	Cluster sampling	43
	Opportunity sampling	45
	Stage sampling	45
3	SAMPLE SIZE	46
	Size versus representativeness	46
	Constraints	47
	Non-response	47
4	CONCLUSIONS	48
	Sampling strategy	50
5	INTRODUCTION TO SAMPLING STATISTICS	51
	Means and standard deviations	52
	The normal distribution	54
	The sampling error	56
	Confidence in sampling estimates	57
	Standard error of a percentage	58
	Hypothesis testing	58
	ACTIVITY	59
	ANSWERS	66
	REFERENCES	68

Objectives

By the time you have completed this part of the block you should be able to:

1 Identify the principles behind random sampling and select situations where random sampling is necessary.

2 Give the reasons why it is unlikely that any sample will be a perfect miniature of the population from which it is drawn.

3 List the advantages and disadvantages of the sampling designs outlined.

4 Determine suitable sample size.

5 Compute the various measures of average, standard deviation, standard error.

6 Recognize a normal curve of distribution.

7 Estimate a population parameter from a sample statistic.

SAMPLING

1 INTRODUCTION

The need for sampling

1.1 In the course of a conversation a short time ago I began talking about Canada. I was stopped in mid-sentence by the person I was talking to with the remark that 'I don't like Canadians – I met one once.' This example might seem absurd, but it highlights the fact that we often make inferences about people, or groups of people, which are based on limited experience and knowledge of them. You might reflect for a moment and ask yourself whether some of your judgements have been of this sort. What this person did, in effect, was to take as a sample of a larger population one member, and on the basis of his findings inferred that because his sample had certain unpleasant characteristics, then so too had the population from which the sample came. This is fine in theory since a fundamental purpose of educational research is to make inferences about the characteristics of populations from the characteristics of samples of those populations. This being the case, why do you think this person was wrong to do so in this instance? Write your ideas in the space below.

1.2 One of the key words in this connection is 'representativeness'. We cannot make any valid generalizations about the population from which a sample is drawn unless the sample is representative of that population. But representative in terms of what? Height? Sex? IQ? Dirtiness? The basic answer is that we want the sample to be representative in terms of those variables which are known to be related to the characteristic(s) we wish to study. Measures of such population characteristics are called *parameters*. (Measures of these characteristics obtained from samples are called *statistics*.) In some cases the variables related to the population parameters come to light in the review of the literature, but in others we do not know what they are. This gives rise to the adoption of different sampling designs. *Stratified* sampling would be more appropriate in the former case, and *simple random*, or *systematic*, sampling in the latter. These terms will be explained later in the block. In answering the question in paragraph 1.1, many of you will have no doubt mentioned the small size of the sample. Pretty obviously we are on dangerous ground to generalize from a sample of one. Besides commonsense, however, there are sound technical reasons why such a sample is inadequate. As we shall see later the smaller the sample, the lower the accuracy.

1.3 But size without representativeness is *never* enough. Perhaps the most infamous example of this point is the poll on voting intentions carried out by the American journal *Literary Digest* in 1936. Although a sample of ten million was used, their forecast of the outcome of the US presidential election was disastrously wrong because the list was picked from telephone directories, thus

biasing the sample in favour of the middle class who were predominantly supporters of the Republican Party. To summarize, estimates of any population parameter are likely to have greater error if based on small and/or unrepresentative samples.

1.4 The term 'population' has been used frequently in the preceding paragraphs. Do you think it has been used in the strictly geographical sense? You may have got that impression because of the nature of the example. In the statistical sense, as used here, 'population' is an arbitrary but clearly defined group (people or objects) in which the researcher is interested. It may be any size from one to infinity – e.g. the population of people who are president of the United States, the population of all real numbers. The statistical and geographical definitions of population are similar in the sense that both involve fixing limits, but in educational enquiries the population is defined more specifically. Examples are 'all thirteen-year-old pupils attending state secondary schools in Sheffield'; 'all sixth-form pupils studying A-level Nuffield Chemistry'; 'all four-year-old children attending nursery schools'. The researcher chooses the population, dependent on the nature of his enquiry. Thus in the example concerning personality and attainment in Part 1 of the block, the population would have been defined as 'all eleven-year-old children'. This population must be defined with care however. Using this population as stated would entail the inclusion of children from all schools in all sectors of education, including private as well as special or ESN schools. If the researcher was not interested in generalizing his results to such sectors then the population would be more correctly defined as 'all eleven-year-old children in state primary schools'. The results gained for this population could not then be generalized to private or special school populations. The relationships between personality factors and attainment could be very different in these latter schools. If, on the other hand, the researcher decided to include these schools and then found that permission to acquire a sample would not be granted, the population would have to be redefined to exclude them. This highlights the fact that occasionally sampling problems define the population, rather than the other way round.

1.5 The children you will hear reading in the radio programme associated with this block were part of a sample from a population of eight-year-old children in a certain London borough. Do the reading standards of the sample enable us to estimate the reading standards in another London borough? In Birmingham? Nationally? The answer in each case is no. Strictly speaking we cannot generalize the results from any *one* sample beyond the population from which it was drawn. What we have to do is wait for comparable data from a number of samples before we can generalize to any wider population. Accumulating evidence from different samples can increase our *confidence* in extending the results beyond the one study we have carried out ourselves.

The random sample

1.6 Once the population has been carefully defined a representative sample can be drawn. The first step towards representativeness is achieved by *random* sampling. Random sampling is that method of drawing a sample so that each member of the population has an equal chance of being selected. This will become clear by the following example. Suppose the population in which we are interested was the first year group of a four-form entry secondary school. Suppose also that the parameter in which we were interested was the level of arithmetic attainment, but of the total population of 99 children we could only test one-third, i.e. 33. Since the population is clearly defined it can, in principle, be listed, i.e. a list including each child's name or code

number. One method of drawing the sample often used is to transfer the list, or 'sampling frame' as it is sometimes called, onto slips of paper. These are then shuffled in a drum. The slips are drawn at random from the drum and the name or number noted.

1.7 At this point there are two alternatives open to us. We can replace the slip in the drum, in which case, of course, when we continue the sampling process the same child may be drawn more than once (*sampling with replacement*). On the other hand we can retain the slip and continue drawing slips from those remaining in the drum until the required sample size, 33, is obtained. There are good statistical grounds for preferring the latter procedure (Moser and Kalton, 1971). The sample obtained by either method is called a *simple random sample*. Its most important feature is that *all* members of the population have *an equal* chance of being selected for it.

1.8 The drum technique would obviously be a laborious method of drawing large samples. An alternative is to use a table of random numbers. These consist of blocks of digits in random order as illustrated in Table 1. (You could generate your own table of random numbers by simply marking ten cards with the digits 0 to 9, and drawing them randomly by the drum technique.) To draw a random sample from our population by this technique, each member of the population would be given a two digit number, i.e. from 00 to 99. The table can be entered at any point, and may be worked through by rows or by columns. If we started in the top left hand column, then the first member of the sample chosen would be number 20, followed by 74, 94 and so on. Alternatively, if we worked by rows the first number selected would again be 20, followed by 17, 42 and so on.

Table 1 Table of random numbers

2017	4228	2317	5966	3861	0210	8610	5155	9252	4425
7449	0449	0304	1033	5370	1154	4863	9460	9449	5738
9470	4931	3867	2342	2965	4088	7871	3718	4864	0657
2215	7815	6984	3252	3254	1512	5402	0137	3837	1293
9329	1218	2730	3055	9187	5057	5851	4936	1253	9640
4504	7797	3614	9945	5295	6985	0383	5187	8556	2237
4491	9949	8939	9460	4849	0677	6472	5926	0851	2557
1623	9102	1996	4759	8965	2784	3092	6337	2624	2366
0450	6504	6565	8242	7051	5501	6147	8883	9934	8237
3270	1772	0361	6626	2471	2277	8833	1778	0892	7349
0364	5907	4295	8139	0641	2081	9234	5190	3908	2142
6249	0090	6786	9348	3183	1907	6768	4903	2747	5203
6100	9586	9836	1403	4888	5107	3340	0686	3376	6857
8903	9049	2874	2104	0996	6045	2203	5280	0179	3381
0172	3385	5240	6007	0671	8927	1429	5524	8579	3196

If the required sample size was between 100 and 999 then three digit numbers could be assigned. Selection also takes place in three digits. Working by rows the first numbers selected would be 201, 742, 282 etc. Numbers occurring outside the population range are disregarded. Note that the sample of the primary school survey, quoted in Chapter 3 of *ERIA* (which you will be reading later) was drawn in this way.

1.9 From data collected from a random sample it is possible to make *inferences* about the characteristics (parameters) of the population from which it was drawn. In fact such inferences can be made only if the sample has been drawn randomly. *The random sample permits us to generalize our research results to the wider population from which our sample comes.* But notice the point that the generalization is an *inference*. Even if the sample is drawn randomly, it is never a completely accurate reflection of the population from which it was drawn, since not every member of the population participates. There is always some

sampling error. In *estimating* the population parameters from the sample we must therefore rely on *statistical inference* and *probability*, rather than certainty. You will remember we looked at *scientific inference* in Part 1. Any generalizations we make are *never certain, only probable*. We shall see how to estimate probability later.

1.10 The following exercise, using the table of random numbers, is designed to demonstrate that errors occur in random samples. In Table 2 there are ten columns. In each column, list ten single digit numbers drawn at random from Table 1. (The first column is done for you.)

1.11 To draw these random samples write the numbers 1 to 15 on separate slips of paper representing the *rows* of Table 1 and 1 to 10 on separate slips of paper representing the *columns* of Table 1. Place each set of slips in a separate box and shuffle them around. Draw out one slip from the first box and one from the second. The number of the first slip locates the row for your first random number and the second slip locates its column. Take the first number in the set of four in the table you have located. Thus if the numbers you draw first are 15 and 2, your first random number will be 3. After drawing each slip replace it in its box. Continue the drawing process until you have selected all the random numbers you need.

1.12 When you have drawn the remaining nine samples of ten random numbers, compute the arithmetic mean (average) for each column. This is computed by adding all the numbers and dividing this total by the number of numbers in each column, in this instance, ten. The formula for this is $M=\frac{\Sigma X}{N}$ where X stands for the value of each individual number, N the number of cases and Σ is a summation sign signifying 'the sum of all the Xs'. The formula therefore reads 'add all the numbers together and divide by the number of cases'. The mean of the first column has also been computed for you.

Table 2 Samples of random numbers

	1	2	3	4	5	6	7	8	9	10
	9									
	7									
	6									
	7									
	3									
	8									
	4									
	1									
	3									
	2									
Total	50									
Mean	$\frac{50}{10}=5.0$									

1.13 In this instance we know the population parameter. The population is the total set of numbers in the table and, as each one from 0 to 9 occurs an equal number of times, the mean must be the mean of 10 numbers 0–9

$$\text{i.e. } \frac{0+1+2+3+4+5+6+7+8+9}{10}=4.50$$

The mean is therefore 4.5. We can now assess the error in the ten random samples. Do any of your sample means exactly equal 4.5? It is not likely. In a practice run of this exercise my sample means varied between 2.6 and 5.9 and 4.6 was the sample mean nearest the population mean. In other words, estimates of population parameters from random samples of those populations are never completely accurate; there is always error. The value of random sampling is that we can estimate, by statistical methods, the likely amount of error. We shall see how later.

1.14 Now compute the mean of the sample means by adding all ten mean scores together, and dividing by ten (the number of samples). You should find that this mean is a more accurate estimate of the population mean than were most of the sample means. This illustrates another important point, mentioned earlier: *in general, the larger the sample the smaller the sampling error.*

2 SAMPLING DESIGNS

2.1 The *simple random sampling design* has just been considered. Other methods of drawing samples exist, the choice of which depends on such factors as the purpose of the study and availability of samples.

Systematic sampling

2.2 If the defined population can be listed – i.e. a 'sampling frame' can be obtained from it – then the samples are sometimes drawn from fixed intervals on the list or frame. The interval is decided by the ratio of the sample population to the total population. In an earlier example, the sample chosen was 33 first year secondary school children from a total defined population of 99. In this instance the *sampling ratio* would be 33 in 99, or 1 in 3. A starting number between 1 and 3 is then chosen at random, and selection continues thereafter at the interval determined by the sampling ratio, throughout the whole list. Thus, if number 2 had been chosen as the starting point, the sampling ratio of 1 in 3 would determine that numbers 5, 8, 11, 14, 17 etc. would be the successive numbers selected. The selection would continue until the required sample size had been acquired.

2.3 If the names in the sampling frame are randomized initially, this method is equivalent to the simple random sampling technique. But in many instances the frame could be arranged in ways that would affect the representativeness of the sample chosen. If the sampling frame consisted of children in a comprehensive school arranged in order of age, and two samples of ratio 1 in 10 were drawn, one commencing 1, 11, 21 and the other 10, 20, 30, then a systematic difference would occur in average age, and possibly therefore in other aspects. The first sample would be older than the latter by nine ranks since child 10 will be younger than child 1, child 20 younger than child 11, and so on.

2.4 The very fact that a sampling interval is used demonstrates the fundamental difference between systematic and random sampling. One of the conditions of random sampling is that the selection of one individual should be independent of the selection of another. This condition is not satisfied in systematic

sampling. A sampling interval of, for example, 1 :4 starting at number 3 ensures that number 7 is certain to be chosen but 4, 5 and 6 have no chance of being selected. However, if the frame is ordered randomly, systematic and random sampling can be considered to be equivalent. The large-scale study *Children and their Primary Schools* used systematic sampling as you will see when reading Chapter 3 of *ERIA*.

Stratified sampling

2.5 In this design the researcher divides his population into layers or *strata*. The characteristics which define these strata are usually those variables which are known *to relate* to the characteristic(s) under study. Thus a population can be divided on the basis of social class membership, sex, level of intelligence, type of school attended if these are known to be important. Having stratified the population, a simple random, or systematic, sample is drawn from each stratum.

2.6 The number of subjects drawn from each stratum will depend on whether the sampling is done proportionately or disproportionately. *Proportionate* stratified sampling requires that the proportion of subjects in the sample should reflect the proportion of subjects in the population. Suppose the population was defined as all secondary school children in one county, and this population was divided up, or stratified, by type of school. If 5 per cent attend public schools, 20 per cent grammar schools, 25 per cent comprehensive schools and 50 per cent secondary modern schools, then these are the proportions of the population that should appear in the sample.

2.7 To draw such a sample in theory would involve listing the whole population of secondary school children by type of school attended, then drawing the sample by simple random or systematic techniques from each stratum until the required numbers were gained. If either systematic or simple random sampling was used, the sampling ratio would be the same for each stratum for the sample to contain the correct proportions of children in each type of school. The following example will illustrate this. Suppose, for the sake of argument, the total population of secondary school children in the county was 100,000, and that because of limitations of time and finance etc. the sample was limited to 1000. This indicates a sampling ratio of 1 in 100 for each stratum. The number of children attending each type of school in the population, and in the sample, is shown in Table 3. Since a *proportionate stratified sample* reflects the actual proportions in the population, greater representativeness is achieved than in a *simple random sample*. Furthermore it can be shown that provided the principle of stratification is *not* completely irrelevant to the study, e.g. type of school bears no relation to the parameter we wish to investigate, then the error in the estimate of this parameter will be less than in a non-stratified sample (Moser and Kalton, 1971). But against this advantage must be set the disadvantage that the size of the sample in such strata can still be too small for adequate analysis on their own. Note that in our example the public school sample is only 50. This means that we would still be severely limited in investigating the characteristics of public school children separately. Stratification with proportionate sampling increases the representativeness of the total sample. But the total sample size still imposes severe limitations on any one stratum that may be involved in it.

Table 8 Numbers in population and sample in a proportionate sampling design

Type of school		Per cent in population	Total in population	Total in sample
Public	X_1	5	5000	50
Grammar	X_2	20	20,000	200
Comprehensive	X_3	25	25,000	250
Modern	X_4	50	50,000	500
Total		100	100,000	1000

2.8 One method of overcoming this is to adopt a *disproportionate* sampling design. This is often used when

(a) proportionate representation makes the sample of one stratum too small for adequate analysis, and

(b) the researcher is particularly interested in one stratum and wishes it to be over-represented.

This is achieved by using a different sampling ratio for the stratum in question. To increase the public school sample the sampling ratio could be altered to 1 in 50 thereby doubling the sample size. To maintain the total sample size at 1000, the sampling ratio of one, or all, of the other strata would also have to be adjusted.

2.9 However, it would no longer be possible to add the data together from the different strata to obtain, for example, the mean for the whole sample. We have to 'weight' the data to compensate for the disproportion that has been introduced. This is done by multiplying the mean score of the public school sample by the proportion of public school children in the total population and similarly for the mean scores in the other strata. These weighted means are then added together. For example, if the mean attainment scores obtained from the four strata are 16 (public), 22 (grammar), 16 (comprehensive) and 14 (modern), then the weighted mean for the total sample is

$$\left(16 \times \frac{5}{100}\right) + \left(22 \times \frac{20}{100}\right) + \left(16 \times \frac{25}{100}\right) + \left(14 \times \frac{50}{100}\right) = 16.2$$

Cluster sampling

2.10 In educational research there is often a gap between sampling theory and practice, because the designs outlined above are often unrealistic to use. If a researcher is using, for example, all thirteen-year-old grammar school children in the country for his population, then theoretically this would involve listing these before selecting on a random basis – a mammoth if not impossible task. The likely outcome would be that he would be using only one or two children in each school. These difficulties are increased when one considers other constraints such as time, finance and administration. One alternative is to select 'natural groups', e.g. schools or classes of children, at random. Sampling of entire groups rather than individuals in this manner is called 'cluster sampling'. Although it has certain theoretical disadvantages, as we shall see later, it retains the principle of randomness and allows a research design within the scope of the individual researcher. Furthermore, lists of the whole population are not required – only for the selected clusters.

2.11 The efficiency of cluster sampling depends on the number and size of the clusters used. At one extreme a large number of small groups could approach simple random sampling: in fact the report on 'Early Leaving' (Chapter 3, *ERIA*) used a derivative of this design. At the other extreme taking one large group such as a school at random, although more convenient, is likely to provide an unsatisfactory sample because of increased sampling error. The error is greatest when the clusters are large and homogeneous with respect to the variable under study. The term 'homogeneous' in this context refers to groups of whom the members are very similar with respect to the characteristics one is studying. The radio programme 'Why Sample?' makes this point by demonstrating the dangers of taking the level of reading attainment in any one school as representative of the reading level for a whole borough.

2.12 Primary schools are comprehensive in the sense that no academic selection procedure operates, so that, theoretically, each school can have an all-ability intake. In practice they have a limited intake area, often restricted to the immediate neighbourhood. Thus the school population tends to reflect the neighbourhood population in terms of social class membership. Suburban schools therefore tend to be 'over-populated' with middle-class children, and schools nearer the town or city centres with working-class children. In general, middle-class children score higher on intelligence tests and show higher standards of reading than working-class children. The classes in such schools will therefore tend to be homogeneous with respect to reading standards – the former will tend to have high reading standards; the latter will tend to have relatively low reading standards. Homogeneous groups are not therefore representative of the total population from which they are drawn; only those parts of it which are homogeneous in the same way.

2.13 If a few large homogeneous groups were adopted as the sample, what would be the likely amount of error in our estimate of the population parameter? To answer this question let us take an example based on figure 2 of your 'Why Sample?' broadcast notes. Presume the study is designed to assess the number of children with a reading problem, the population defined as 'all children attending state primary schools in the London borough of Waltham Forest'. A 10 per cent sample has been decided upon, which in this case is approximately 300 children, and to save time and travel etc., the researcher decides to sample whole schools. Even by random sampling procedures, schools 1, 16 and 28 (counting from the left) could have been selected, which gives approximately the required sample size. What do you notice about these three schools? All are homogeneous with respect to the variable under study (i.e. they all have only slight reading problems). Are they likely to give an accurate estimate of the reading problems in the whole population? We can check this because in this instance we know the population parameter. The number of children with a reading problem in the population is 786. The The number with a reading problem in the sample is found by:

$$\left(\frac{15}{100} \times 122\right) + \left(\frac{8.5}{100} \times 82\right) + \left(\frac{10}{100} \times 155\right) = 41$$

SELF-ASSESSED QUESTION 1

Comparing these figures, do you think it is likely that the sample estimate is an accurate estimate of the population parameter? Can you suggest a better method of cluster sampling for this example? Write your answer in the space provided on p. 45.

Turn to p. 66 for a model answer.

Butcher's (1966) advice is to err on the safe side: assume a fairly high degree of homogeneity within schools and classes, and spread the net as wide as possible.

Opportunity sampling

2.14 A derivative of cluster sampling which involves even greater error is where the researcher deliberately picks the elements in the population he wishes to study. *Deliberate* or *opportunity* sampling (sometimes also called *quota* or *judgement* sampling) is often used in educational research because there is no other alternative open to the researcher. This may be the case when one can afford to sacrifice a certain amount of accuracy such as in a pilot enquiry, or when he wants to study minority groups for whom no sampling frame is available. For example, if he wanted to compare schools that use 'informal' and 'formal' teaching methods he would probably have to rely on experts to select such schools, on the basis of their personal knowledge, for each category. (This was done by Haddon and Lytton in the study referred to in Part 1.) There is no general listing of schools that can be classified in this way.

2.15 Deliberate sampling may produce biased samples, and therefore greater likelihood of error. The trouble is we can never know for certain because strictly speaking it is not possible to estimate the sampling error in such samples, though it is often done on the very weak assumption that any sample is one of the *possible* random samples that could have been selected. Opportunity samples are thus valuable in exploratory studies, but clearly the results cannot be generalized to any wider population. They should therefore be used only if the elements of the population of interest in the study cannot be found in any other way. It may be of interest to note that much 'opinion polling' is based on opportunity samples. This may help to explain why so much apparently contradictory evidence is collected by the market research agencies who carry it out.

Stage sampling

2.16 A further alternative exists which falls midway between the random sample and its derivatives, and cluster sampling. That is, it avoids the often impossible rigour of the former, but ensures wider representation than the latter. This is *stage* sampling. A two-stage design would first take a random selection of schools, and within each school a random sample of children. A two-stage design was used by the Primary School Survey which studied the necessity for auxiliary help in such schools. This study was concerned with classes rather than children, and therefore they first selected schools at random, and then selected classes within these schools on a random basis. For larger studies multi-stage designs can be introduced – regions within the country would be randomly selected; within these, local education authorities (LEAs) would be randomly selected; within LEAs, schools would be randomly selected; within schools classes would be selected; within the classes children would be selected, and so on. Such designs obviously lie beyond the resources of most individual researchers. But their great strength is that they permit accurate

estimates of the sampling error. This is because at each stage the principle of *random selection* is maintained.

2.17 A number of large scale studies in education have used the different sampling designs which we have considered in this section. Some of these are reviewed in Chapter 3 in *ERIA*. Read this chapter and note in the space provided below the type of sampling design used in each study.

Study	Sample design

3 SAMPLE SIZE

Size versus representativeness

3.1 In general the larger the sample the better, simply because a large sample tends to have less error, as we found in the exercise using the table of random numbers. The radio programme 'Why Sample?' makes the same point. This is not to say that a large sample is sufficient to guarantee accuracy of results. Although for a given design, an increase in sample size increases accuracy, it will not eliminate or reduce any bias in the selection procedure. We have already seen an example of this principle in the *Literary Digest* survey mentioned at the beginning of this part of the block. As another example, if a 1 in 2 sample of the whole country consisted of one sex only it would be large, but unrepresentative. Size is therefore less important than representativeness. Fox (1969) put forward a general adage for research on this basis: 'No data are sounder than the representativeness of the sample from which they were obtained, no matter how large the sample!'

3.2 Another consideration is whether sub-groups are to be analysed in the study, and if so, to what extent. Many educational research projects not only make general comparisons between the major groups studied, but gain valuable information by dividing these into sub-groups, for comparison. In the stratified sampling example (paragraph 2.6), sub-groups were formed on the basis of type of school. They could have been divided again on the basis of sex. At each sub-division, the numbers in the sub-groups – public school boys, grammar school girls – become smaller, and at a certain level begin to create problems for statistical analysis. The sample could be divided in many ways dependent on the measures used. If a test of intelligence had been administered to the sample in the above example, other sub-groups defined by this characteristic could have been created. This demonstrates the necessity of

disproportionate sampling if the researcher is interested in studying a particular sub-group. In fact many sampling designs start from considerations of how large certain sub-groups in the sample are to be, and then go on to consider how such sub-groups are to be obtained by random sampling methods. As noted in Part 1 of the block many questions of this kind may be answered by an adequate pilot investigation before the main study is carried out.

Constraints

3.3 Finance, time, facilities and amount of assistance for collecting and analysing the data are all crucial in deciding sample size. The manner in which the number of classrooms was chosen for the primary school survey epitomizes this (Duthrie, 1970, in Chapter 3 of *ERIA*):

> The approximate number of classes which we could observe was thus dictated by three factors; the number of pairs of observers available (two); the amount of time available (six months); and the length of time to be spent in each class (one day). Bringing these three factors together and allowing time for travel and for the analysis of the data, it was possible for the Team to observe, in the course of the main study, approximately 120 classes.

Other constraints also operate. The size of the sample can be reduced through non-co-operation on the part of the schools selected, or by refusal of directors of education to allow research in schools under their local education authorities. The radio programme associated with Block 1 examined these constraints in more detail.

Non-response

3.4 In deciding size, account should also be taken of possible attrition of numbers; even the best designed projects lose about 10 per cent of their sample, depending on the nature of the study. This is particularly important in follow-up studies in which the same children are retested at a later date. There is always the likelihood that children have left the area between visits, but of equal importance is the level of school absenteeism. Manchester was recently reported (BBC news programme) to have a 25 per cent absence rate within one week. Losses of this order could seriously affect the representativeness of the sample and cause increased error. The children who stay away may have quite different characteristics from those who attend. For example, in any study of children with learning problems this group is an important section of the potential sample. Studies which make use of postal questionnaires are also particularly prone to sample loss through non-response. Initial returns are often as low as 50 or 60 per cent, although there are means of securing larger returns, such as follow-up letters. Unless returns in excess of 80 per cent can be gained, bias due to non-response can seriously affect the representativeness of the sample. Efforts are often made to locate and interview random samples of non-respondents to ascertain whether their characteristics differ from those of respondents sufficiently to affect the results. If the reasons for non-response are unconnected with the purposes of the survey, then they need not increase bias. If they are connected, this leads to increased error and casts doubts upon the generalizability of the results. Note that the response rate in the *Literary Digest* study was only 20 per cent.

3.5 One final point is worth mentioning in relation to sample size. There is no point in wasting resources of time, money and personnel to obtain a sample that is larger than is needed to carry out the analysis one wants to do. All the factors limiting the size of the *obtained sample* need to be taken into account when deciding this question. We need a *sufficient sample* to meet the analysis needs of the investigation we are undertaking.

4 CONCLUSIONS

4.1 Completing the following table will provide a useful summary of the sample designs which have been considered in the previous two sections. The type of sample is indicated in column 1. You should complete column 2, describing how each sample is drawn, and column 3, the advantages and disadvantages of each method. In completing column 3 keep in mind such things as representativeness, error, appropriateness and/or convenience to the researcher.

Table 4 Summary of research designs

Column 1	Column 2	Column 3
Type of sample	How sample is drawn	Advantages and disadvantages
1 Simple Random		
2 Systematic		
3 Stratified (a) Proportionate		

4 Stratified (b) Disproportionate		
5 Cluster		
6 Deliberate		
7 Stage		

4.2 In this part of the block a recurring theme has been that theory and practice in sampling in educational research are two different matters. The researcher constantly finds in the practical situation that any neat, theoretical sampling plan has to be modified. A choice has to be made between a plan that is tidy and will permit straightforward analysis but will be less practicable in terms of administration, and one that is less tidy theoretically but is within the scope of the researcher. These are no doubt some of the reasons why, in educational research, the use of random samples are, according to Butcher (1966) 'as common as white blackbirds'.

Sampling strategy

4.3 Butcher has sketched out a rough sequence for sample planning which outlines the alternatives, although the proviso is made that it is a set of suggestions, not of rules; a general guide, not a blueprint.

1 Define clearly the population about which you intend to generalize from your results.

2 Estimate roughly the size of sample required.

3 Consider the possibility of a simple random type sample. In most studies this will not be practicable, but if doing something like a postal survey on parental attitudes in a certain area and a complete sampling frame is available then it could be considered.

4 If simple random sampling is not practicable decide if stratification would increase convenience or efficiency. For example, if testing attainment the sample might profitably be broken down by school type, or social class.

5 If not 4 then consider using clusters selected on a random basis. Use several small clusters rather than a few large ones, or preferably a two-stage or multi-stage design which will cover a wider range of clusters for the same sample size.

6 If 5 is not possible then the use of non-random samples must be considered. Such samples cannot be decried altogether but the crucial problem is that it is not possible to make, from the results themselves, a theoretically satisfactory estimate of the sampling error which is always present.

7 Even an opportunity sample may yield information of considerable value – particularly in an exploratory study of an area previously under-researched or in 'case' or 'clinical' studies using a small number of individuals. The value of any conclusion from these types of study will depend on an intelligent assessment of how far the conditions have been satisfied for valid scientific generalization. This is also an important point to take into consideration when reviewing research literature. McNemar writing in 1946 made the point: 'The existing science of human behaviour is largely the science of the behaviour of sophomores.' When you read psychological research journals, you may think his criticism is only slightly less valid today.

WE CANNOT MAKE ANY VALID GENERALIZATIONS ABOUT THE POPULATION FROM WHICH A SAMPLE IS DRAWN UNLESS THE SAMPLE IS REPRESENTATIVE OF THAT POPULATION.
school
REPRESENTATIVENESS IS ASSUMED TO BE ACHIEVED BY RANDOM SAMPLING.
RESEAR
RANDOM SAMPLING IS THAT METHOD OF DRAWING A SAMPLE SO THAT EACH MEMBER OF THE POPULATION HAS AN EQUAL CHANCE OF BEING SELECTED
RESEARCH
ESTIMATES OF POPULATION PARAMETERS FROM RANDOM SAMPLES OF THOSE POPULATIONS ARE NEVER COMPLETELY ACCURATE;
RESEARCH
THERE IS ALWAYS ERROR!
RESEARCH

5 INTRODUCTION TO SAMPLING STATISTICS

5.1 In Section 1 the point was made that when a sample of a population has been drawn at random it is possible to estimate the parameters of the population. It was also stated that in estimating such parameters we must rely on statistical inference and probability. But first we must grapple with some elementary statistical concepts. In introducing these, use will be made of Chapter 5 in the third edition of Thorndike and Hagen (1969). Students already conversant with statistics should also read Chapter 4 in Guilford (1965).

Means and standard deviations

5.2 Before continuing, read pages 133-46 in Thorndike and Hagen and work out the following:

SELF-ASSESSED QUESTION 2

(a) Complete a grouped data frequency distribution, a histogram, and a frequency polygon (as you learned to do in the radio programme) of the spelling scores in Table 5.1 of the above book.

(b) What is the value of

the mean? _____

the median? _____

the mode? _____

The answers to (b) are at the end of the block, p. 66.

5·3 In sampling, the average or measure of a central tendency, as expressed by the *mean*, the *median* or the *mode*, is an important statistic, because we use the average of a randomly drawn sample as an estimate of the average in the population. The mean is the measure usually adopted because it is a more stable measure, and because other measures can be calculated from it which we need to estimate the error in the estimate of population mean. However, the mean, although a useful description of a sample, gives us a very limited picture of it. For example, two classes may have gained the same average score for reading ability, but these averages could come about in different ways. One class could have had a large number of very good and a large number of very poor readers; whereas the other could have had a large number of children *all* of whom were *neither* very good nor very poor readers. The frequency polygon of these two classes would therefore have been quite different. Can you visualize what shapes these polygons would take? The first would look something like a camel's back, whereas the second would be narrower with a sharp central peak. In other words the dispersion or variability of the two sets of scores would be different, i.e. the amount of spread between high and low scores. One measure of this variability is the *range* of scores, which is simply the difference between the highest and lowest scores.

SELF-ASSESSED QUESTION 3

Which of the two classes had the greater range? (*The answer is on p. 66.*)

5·4 There are several measures of variability as you will see if you read pages 146-50 in Thorndike and Hagen, but we will concentrate on the *standard deviation* since this is derived from the mean. To demonstrate how this is computed we will use as an example a small group of ten children who have been given an English test. The scores gained by each child are shown in Table 5:

SELF-ASSESSED QUESTION 4

Table 5 Distribution of English scores

Pupil	Score (X)	Deviation (x)	Squared Deviation (x^2)
1	22	7	49.0
2	12		
3	14		
4	12		
5	14		
6	14		
7	17		
8	19		
9	11		
10	15		
	Mean (M) =		Σx^2 =

5.5 The formula for computing the standard deviation (σ)* is

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

where x^2 equals the sum of the squared deviations from the mean, and N equals the number of cases in the sample. In other words it is the square root of the average of the squared deviations from the mean. You will find this easier to understand once we have computed it for the set of scores above. It requires six steps. Carry out each step on the data in Table 5 before progressing to the next step.

1 Compute the mean.

2 We next need to know how much each score deviated from the mean. This is found by subtracting the mean from each score ($x = X - M$) then placing the result in the 'deviation' column. The first one has been done for you. You will notice when you do this yourself that some of these deviations are *negative* and some *positive*. A check on your computation can be used here since the sum of all the deviations should be equal to zero.

3 We wish to make use of the deviations, whether *positive* or *negative*, so we square each deviation and put the results in the 'squared deviation' column. Again the first one has been done for you.

4 Add all the squared deviations together. This gives the sum of the squared deviations (Σx^2).

5 To find the average of the squared deviations, divide by the number of cases in the sample (N), i.e. $\frac{\Sigma x^2}{N}$

6 Having found the value of $\frac{\Sigma x^2}{N}$, find the square root of this number, i.e., $\sqrt{\frac{\Sigma x^2}{N}}$. The square root can be computed by hand as in Thorndike and Hagen, p. 657, or via the statistical tables in Guilford, Table A, Appendix B. The standard deviation of the scores in Table 5 is 3.26, correct to two decimal places. If your answer differs from this, turn to p. 66 at the end of the block, where you will find the complete table.

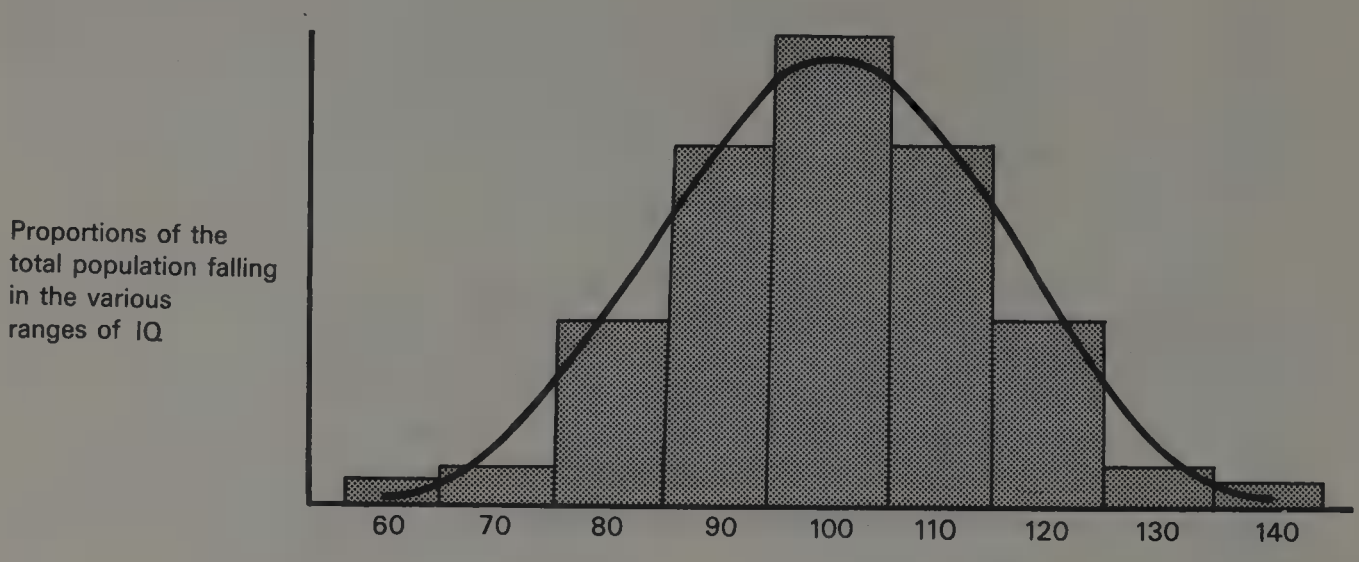
The normal distribution

5.6 The value of the standard deviation will become apparent when we have examined the normal distribution curve. The normal curve has a symmetrical, bell-like shape as illustrated in Fig. 1, which shows the proportions of the total population falling into various ranges of intelligence. The mean IQ is 100 and you will see that many of the cases cluster round the mean. This simply indicates that many people have average levels of IQ, less people have high or low levels, and few have very high or very low levels of IQ. Many other human characteristics are similarly distributed. Most of us are average height; fewer are taller than 6 ft. or shorter than 5 ft. 3 in., and even fewer are taller than 6 ft. 6 in. or shorter than, say, 4 ft. 9 in. Therefore if we construct a frequency polygon of the population in terms of height, the resultant curve closely conforms to the normal curve. If you examine the curve of the spelling scores from Thorndike and Hagen, you will see that it is not completely symmetrical but has a 'jagged' profile. This is because the sample was insufficiently large to

*It is a convention in statistics that Greek letters are used for the population parameters of which sample statistics are estimates. However, to conform with the notation that Guilford uses, we are using the Greek letter σ (sigma) to represent the sample value of a standard deviation and $\bar{\sigma}$ to represent the population value.

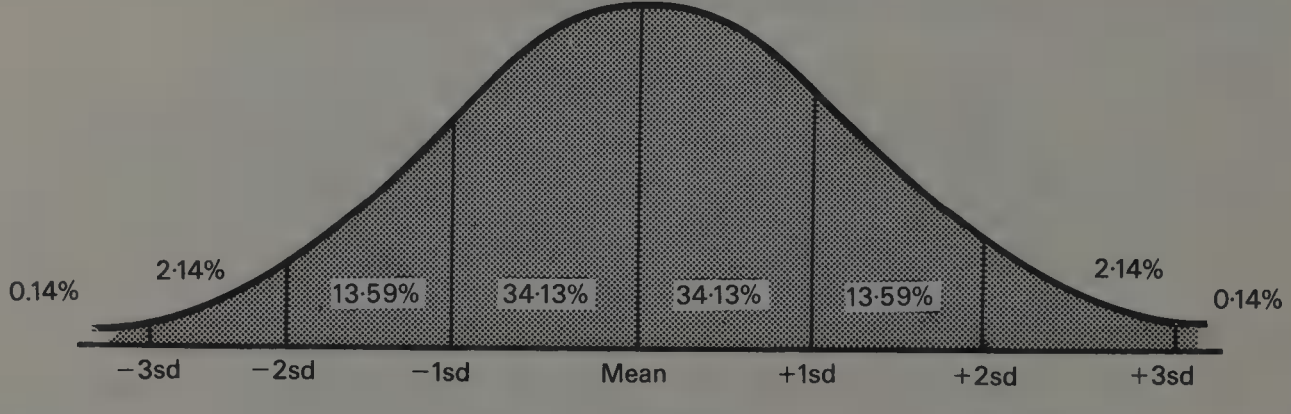
ensure a normal curve. The characteristics of a small sample are therefore less likely to be normally distributed even though they may approximate to a normal distribution.

Fig. 1 Diagram of the normal curve fitted to IQ data



5·7 Note also in Fig. 1 that the frequencies of the various scores (expressed as proportions of the total number) are represented by the areas of the vertical bars in the histogram. Fig. 2 shows the normal curve with the mean marked in the centre of the baseline, and the standard deviations (sd) marked off in both directions above and below the mean.

Fig. 2 Proportion of normal curve with perpendiculars erected both at the mean and at each standard deviation above and below the mean.



When perpendiculars are drawn from these deviations, slicing the distribution into bands one standard deviation wide, a fixed percentage of the cases always falls in each band. This indicates that in a normally distributed set of scores, a little over two-thirds (68.26 per cent) of all the scores lie within plus or minus one standard deviation from the mean, 95.44 per cent lie within two standard deviations and so on.

SELF-ASSESSED QUESTION 5

What percentage of scores lies between

(a) plus and minus two standard deviations? _____

(b) plus and minus three standard deviations? _____

See p. 67 for the answer.

5.8 For practical purposes the full range of a normal distribution is just over *six* standard deviations (i.e. 99.72 per cent of normally distributed observations will fall within this range). This knowledge can be used to provide a convenient first test of whether a set of scores is normally distributed. Simply divide the range by the standard deviation. If the result is approximately six, then the data can be considered normally distributed. If the result is more than six this indicates that the distribution is more *elongated* than normal, and if less than six, more *restricted* than normal.

SELF-ASSESSED QUESTION 6

Using this method, assess whether the data in Table 5 are normally distributed. (*Check your answers against those on p. 67.*)

(a) $\frac{\text{range}}{\text{standard deviation}} =$

(b) This value indicates that the data are

(c) The likely reason for this is

5.9 To revert back to the review of the literature for a moment, you will find that researchers often define the groups in which they are interested in terms of standard deviations. For example, a researcher interested in the characteristics of backward readers would first administer a reading test to his sample, and then might define backward readers as those whose scores were lower than one standard deviation below the mean.

SELF-ASSESSED QUESTION 7

On this criterion backward readers would form the bottom [%] of the distribution, assuming the sample was normally distributed. (*See p. 67 for the answer.*)

5.10 As there is an exact mathematical relationship between the standard deviation and the proportion of cases, the same proportion of cases will always be found within the same standard deviation limits. The importance of this in estimating sampling probabilities will become evident after considering the *sampling distribution.*

5.11 Let us suppose we have a population of 3000 students in a college and the purpose of a survey is to estimate the average age. A random sample of 150 students is selected and their mean age calculated. This sample mean we take to be an estimate of the population mean, i.e. of all 3000 students. However, if we return this 150 back to the population and draw another 150 at random and calculate this sample mean, it is almost certain that the two sample means will differ. We therefore have two different sample means, both estimates of the same population mean. If we continue this process and draw, say, 1000 samples, there would be a spread of sample means akin to the spread of the intelligence test scores in Fig. 1, i.e. a normal distribution. The distribution of sample means, logically called *sampling distribution*, might look something like that in Table 6, which lists the frequencies that each sample mean occurred – e.g. ten samples produced sample means between eighteen and

nineteen, ten samples produced sample means between nineteen and twenty years, and so on.

Table 6 Hypothetical distribution of the mean ages of 3000 students

Mean age of sample	Frequency of occurrence in sample
18–19	10
19–20	10
20–21	30
21–22	100
22–23	190
23–24	290
24–25	240
25–26	70
26–27	50
27–28	10
	Total = 1000
	Mean = 23.5

EXERCISE

Construct a histogram and frequency polygon of these data.

The sampling error

5.12 You will see that the distribution conforms closely to the normal curve. The mean of 23.5 can be taken as a fairly accurate reflection of the population mean, as we found in the example using random numbers earlier in this part of the block. It will also be apparent that any one sample of 150 students could fall anywhere within this sampling distribution, and as such is not likely to be a completely accurate estimate of the population mean. To be able to estimate how accurately one sample mean reflects the population mean we need to know something about the dispersion of all sample means. We can calculate the dispersion or variability of sample means by computing a standard deviation, just as we did for a distribution of scores in one sample. The standard deviation of a distribution of sample means is called the *standard error*, signified σ_M. The formula for computing the standard error is $\bar{\sigma}_M = \frac{\bar{\sigma}}{\bar{N}}$

where $\bar{\sigma}$ = the standard deviation of the population mean and N = the number in our sample. But we have repeatedly stated that we rarely know the population mean; in fact, generally the object of our sampling exercise is to find out what it is. And it is unlikely that we can take a number of samples like the 1000 we have just used to estimate it. So how can we know what the standard deviation of the population mean ($\bar{\sigma}$) is? In fact we have to estimate it using the standard deviation of just one sample. In other words we presume that the variability in the sample is an accurate reflection of the variability of a distribution of random samples, i.e. the variability of the population. The formula for the *estimate* of the standard error is: $\sigma_M = \frac{\sigma}{\sqrt{N-1}}$ where σ_M = the estimated standard error, σ = standard deviation of the sample, and N = the number in the sample.[1] The reason for subtracting 1 from the sample size need not concern us here since it will be dealt with in greater detail in the block on statistical analysis (Block 4).

5.13 The value of the estimate of the standard error, together with our knowledge of the characteristics of the normal curve, allows us to say something about the accuracy of the estimate of the population mean from the sample mean.

Confidence in sampling estimates

5.14 We have stressed that it is never possible to find the exact value of a population mean from a sample mean. But we can specify an interval within which the population mean lies, with a known degree of confidence. The interval is called the *confidence interval*. (It is similar to the idea of 'error band' associated with a single measurement, to be discussed in TV3.) You will recall that the area under the normal curve between ± 1 standard deviation includes approximately 68 per cent of the total area. Since a standard error is actually a standard deviation it has the same relationship to the normal curve. Thus, if a large number of samples of the same size were drawn, as in the student example, we would expect that 68 per cent of those sample means would fall between ± 1 standard error. Stated the other way round we can say with 68 per cent confidence that the population mean will fall between ± 1 standard error, i.e. the odds are 68:32 against the mean falling outside this interval. If we wanted to be more confident of the estimate we could say that we would be 95 per cent confident that the population mean would fall between ± 2 standard errors, or 99 per cent confident that the population mean would fall between ± 2.58 standard errors.

5.15 Let us take an example to illustrate this. Suppose we drew a sample of 150 students and found the mean to be 24 and the standard deviation 6.1. The estimate of the standard error σ_M would be

$$\frac{6.10}{\sqrt{150-1}} = \frac{6.10}{12.21} = 0.5$$

We can therefore say that we are 68 per cent confident that the population mean will lie within ± 1 standard error of the sample mean, i.e. $24 \pm 0.5 = 23.5$ to 24.5 years. Similarly we would be 95 per cent confident that the population mean would lie within ± 2 standard errors of the sample mean, i.e. $24 \pm (2 \times 0.5) = 24 \pm 1.0 = 23$ to 25 years, and so on.

[1]This is assuming that the sample size is more than 50. For samples smaller than this a different calculation is required, which we shall be considering in Block 4.

SELF-ASSESSED QUESTION 3

In the following example:

(a) Compute an estimate of the standard error.

(b) State within which limits the population mean would lie at the 95 per cent, and 99 per cent level of confidence.

Example
A random sample of 290 ten-year-old children were given a test of reading abilities with the intention of estimating the reading ability of all ten-year-old children; the mean was 104, standard deviation 5.67.

(a) Estimate of standard error

(b) 95 per cent confident that population mean lies within

(c) 99 per cent confident that population mean lies within

Check your answers against those on p. 67.

Standard error of a percentage

5.16 To conclude the section on standard error, we shall briefly consider the estimation of it for another important statistic – the *percentage* of the sample with a given characteristic. Supposing that the only data we have about the distribution of a characteristic in a sample is the percentage who have it, e.g. the percentage of children who pass a reading test (P). It can be shown that the standard error of such a percentage approximates to the standard error of the mean of a variable with only two values 0 and 1. This reduces to the following formula $\sigma_P = \sqrt{\frac{P(100-P)}{N}}$ i.e. the standard error of the percentage equals the square root of the percentage of children who pass the test multiplied by the percentage of children who fail the test and divided by the total number of children in the sample. You will be working out this standard error for the data presented in the 'Why Sample?' radio programme accompanying this block.

Hypothesis testing

5.17 Often a researcher is not interested in estimating the population parameter but wishes to know whether two or more groups within the population are significantly different with respect to this parameter (e.g. means or percentages). This is part of the general question of how to test statistical hypotheses and will be taken up in the block on data analysis. It is worth mentioning here, however,

that the confidence levels we have been considering in relation to population estimates are also used as the basis for inferring that a hypothesis is refuted or supported. We talk about the .05 (5 per cent) and .01 (1 per cent) *levels of statistical significance* which are the equivalent of the 95 per cent and 99 per cent confidence levels for population estimates.

5.18 In conclusion, you must keep in mind the fact that we have used the standard error and the normal distribution on which to base our estimates of the population parameter, from the sample statistic. In doing so we are assuming that our sampling procedure provides us with a random sample of observations of the characteristic which is normally distributed.

ACTIVITY Designing samples

On p. 61 a sampling frame is listed. It is the entire population of the first year of an imaginary secondary modern school, measured along five dimensions: sex, IQ, social class of father (I/II = professional and managerial, III = non-manual and skilled manual, IV = semi-skilled manual, V = unskilled manual), number of children in the family and score on some imaginary test of attitude towards school (10 = highly positive, 1 = highly negative).

From this population you can draw samples of different size and by various methods and see how these compare with each other and the whole population. It is rare indeed for a researcher to be able to know the characteristics of the whole population (and, statisticians would maintain, unnecessary); instead he estimates them from his sample and assigns statistical probabilities to his estimates.

This activity should give you some feeling for the extent to which these estimates or extrapolations are reasonable. You should spend about forty-five minutes on the activity, but for those of you who enjoy playing with it there is space left on the scoring sheet for further samples – random samples of various sizes is one possibility, or you could make a more searching analysis of the samples you have already drawn, such as boys and girls separately.

1 Random sampling

Enter the random numbers table (p. 65) at any point, and read off the first twenty two-digit numbers (it will be easiest to let 00 represent 100, otherwise the majority of numbers in the table will be outside the range of your sample). It is possible that the same number will have occurred more than once, in which case you will have to decide whether to use replacement or to continue reading off numbers until you have twenty *different* two-digit numbers. This problem was discussed in paragraph 1.7. Record on the scoring sheet how many of your sample fall into each of the categories given, and calculate the mean attitude score. Enter the random numbers table in a different place and select another sample of twenty. Do the same as before, and then compare your two samples with each other and with the whole population. If your sample means differ greatly from the population can you see any obvious reason for this? (E.g. the sample might be heavily biased in some respect.)

2 Systematic sampling

Select twenty cases by systematic sampling (refer to paragraph 2.2 if you need to refresh your memory about this). Compare the attitude ratings with your random samples and the whole population.

3 Disproportionate stratified sampling

Consider the ways in which the population varies and decide on the aspect(s) which will be most relevant to a fair sampling for attitude to school. You will then be in a position to define each of your strata in terms of one or more of the variables. For instance you might decide that the variable 'number of children in family' would strongly affect the attitude scores. You might therefore classify your population into smaller families (one to three children) and larger families (four or more children). You would then list the population of each stratum and randomly select ten from each.

You could of course decide upon much more complex strata such as low IQ and smaller family, low IQ and larger family, high IQ and smaller family and high IQ and larger family.

When you have selected your strata and sampled them randomly such that the total sample size=20, calculate the mean attitude score and compare it with your previous samples.

4 Proportionate stratified sampling

This time instead of simply taking an equal number of cases from each stratum take numbers which reflect those of the total population. For instance larger families and smaller families are in the ratio of 65:35 in the population. The ratio in the sample would therefore be 13:7. Compare the average attitude score with your previous samples, particularly with the disproportionate sample.

5 Quota sampling

This method is more commonly used in market research than educational research, and is discussed under the heading of *opportunity sampling* in the main text. Strata are determined as before, but instead of listing each stratum and selecting from it at random the researcher knocks on doors or accosts people in the High Street etc. until he has the number of respondents he requires for each stratum. Samples drawn in this way are open to great risk of becoming distorted. As the *Literary Digest* did not take into account that telephone subscribers are predominantly middle-class, so it is easy to forget for instance that the majority of people to be found in the High Street will be housewives. This sampling method requires even more careful consideration of which strata are relevant. (For example, the majority of schools willing to co-operate in research projects may well be enlightened and 'progressive'.)

For your quota sample keep the same proportionate strata you just used. Now select a sample without the use of the random numbers table; just scan the list quickly and select the first pupils you find which conform to each stratum until you have the necessary numbers of each. Again calculate the mean attitude score and compare. Does this sample seem to be as representative of the population as your random samples?

Sampling Frame

Number	Boy/Girl	IQ Score	Social Class of Father	No. Children in Family	Attitude Rating
01	B	105	IV	4	3
02	B	98	III	2	7
03	B	102	III	3	6
04	B	120	IV	1	8
05	B	94	III	2	5
06	B	87	V	6+	1
07	B	101	I/II	1	7
08	B	103	III	2	6
09	B	124	III	3	6
10	B	120	III	5	5
11	B	98	IV	4	4
12	B	97	III	3	5
13	B	102	V	2	4
14	B	110	IV	1	9
15	B	94	III	4	4
16	B	108	V	2	6
17	B	100	III	1	6
18	B	96	V	3	5
19	B	98	IV	2	3
20	B	102	III	4	4
21	B	89	III	2	4
22	B	104	IV	3	6
23	B	110	III	2	3
24	B	119	I/II	2	6
25	B	96	III	3	7
26	B	105	IV	2	7
27	B	104	III	3	4
28	B	93	V	4	3
29	B	98	IV	1	8
30	B	108	III	3	6
31	B	102	III	5	4
32	B	99	V	2	4
33	B	86	III	3	4
34	B	104	III	1	7
35	B	96	IV	2	5
36	B	110	III	3	8
37	B	123	III	4	6
38	B	115	V	1	7
39	B	112	IV	3	5
40	B	97	III	2	5
41	B	93	I/II	6+	8
42	B	98	V	2	3
43	B	107	IV	2	6
44	B	100	IV	2	6
45	B	104	III	3	5
46	B	97	III	2	4
47	B	96	IV	4	2
48	B	101	IV	3	5
49	B	108	III	1	7
50	B	92	III	2	5
51	B	118	III	5	7
52	B	104	III	2	6
53	G	94	III	3	5
54	G	106	III	3	7
55	G	115	III	2	8
56	G	88	V	4	2
57	G	97	IV	2	4
58	G	96	V	6+	4
59	G	105	III	3	6
60	G	107	III	2	7
61	G	107	III	1	8
62	G	103	III	4	7
63	G	98	III	3	5
64	G	112	I/II	1	7
65	G	94	IV	2	5
66	G	97	IV	3	4
67	G	103	IV	1	7
68	G	107	III	2	8
69	G	100	V	3	4
70	G	91	V	5	3
71	G	101	IV	3	5
72	G	105	III	2	7
73	G	89	III	5	3
74	G	106	IV	1	7
75	G	120	III	4	6
76	G	101	IV	2	6
77	G	104	III	3	7
78	G	98	IV	4	5
79	G	103	V	3	4
80	G	85	III	6+	3
81	G	110	III	1	8
82	G	108	IV	2	6
83	G	112	I/II	1	9
84	G	99	III	4	6
85	G	95	IV	3	5
86	G	123	III	5	6
87	G	118	V	2	7
88	G	99	IV	4	3
89	G	95	V	2	6
90	G	102	III	1	8
91	G	103	IV	3	5
92	G	102	V	6+	5
93	G	95	III	2	6
94	G	99	III	3	6
95	G	96	IV	3	4
96	G	105	IV	4	4
97	G	99	III	2	5
98	G	110	I/II	3	9
99	G	98	III	2	7
100	G	97	III	3	4

Attitude Scores

Sex

	1	2	3	4	5	6	7	8	9	10	Row Totals
Boy	1	1	5	10	10	12	8	4	1	0	52
Girl	0	1	4	8	9	9	10	5	2	0	48
Column Totals	1	2	9	18	19	21	18	9	3	0	100

IQ Scores

85–89	1	1	2	2	–	–	–	–	–	–	6
90–94	–	–	2	1	4	–	–	1	–	–	8
95–99	–	1	3	8	8	4	3	1	–	–	28
100–104	–	–	–	6	5	7	5	1	–	–	24
105–109	–	–	1	1	–	5	6	2	–	–	15
110–114	–	–	1	–	1	–	1	2	3	–	8
115–119	–	–	–	–	–	1	3	1	–	–	5
120–124	–	–	–	–	1	4	–	1	–	–	6
Column Totals	1	2	9	18	19	21	18	9	3	0	100

Social Class

I/II	–	–	–	–	–	1	2	1	2	–	6
III	–	–	3	8	9	13	11	6	–	–	50
IV	–	1	3	5	8	5	3	2	1	–	28
V	1	1	3	5	2	2	2	–	–	–	16
Column Totals	1	2	9	18	19	21	18	9	3	0	100

Number of Children in Family

1	–	–	–	–	–	1	7	5	2	–	15
2	–	–	3	5	6	10	6	2	–	–	32
3	–	–	–	7	10	6	3	1	1	–	28
4	–	2	3	4	1	3	1	–	–	–	14
5	–	–	2	1	1	1	1	–	–	–	6
6 and over	1	–	1	1	1	–	–	1	–	–	5
Column Totals	1	2	9	18	19	21	18	9	3	0	100

On the previous page the attitude scores are analysed by their frequency along each of the other dimensions in turn. Although this sampling frame is completely artificial, the trends which can be discerned from these analyses reflect those actually found in this age group. Tables 8 and 9 in *The Plowden Children Four Years Later* by G. F. Peaker (NFER 1971) show girls to have a slightly more positive attitude towards school than boys, for high IQ test scores to be related to positive attitude, and for family size and social class (considered as part of the variable 'parental circumstances' by Peaker) to be related to attitude. Your project data will also furnish you with information about the relationship between sex and attitude and between family size and attitude.

The row and column totals indicate the distribution of each of the variables among the population, and it is these you will need to refer to when comparing your various samples. They are listed again in the first column of the scoring sheet. Finally the averages (means) for each dimension for which this is appropriate are given below, both for the total population and for boys and girls separately.

	Population	Boys	Girls
IQ	104.6	104.9	104.2
Family size	2.79	2.7	2.9
Attitude score	5·5	5·3	5·7

Scoring Sheet

Type of Sample and Size		Whole Population (N = 100)	Random (N = 20)	Random (N = 20)	Systematic (N = 20)	Stratified (N = 20)	Proportional Stratified (N = 20)	Quota or Opportunity (N = 20)					
Sex	B	52											
	G	48											
IQ Score	85–89	6											
	90–94	8											
	95–99	28											
	100–104	24											
	105–109	14											
	110–114	8											
	115–119	5											
	120–129	6											
Social Class	I/II	6											
	III	50											
	IV	28											
	V	16											
Number of Children in Family	1	15											
	2	32											
	3	28											
	4	14											
	5	6											
	6 and more	5											
Attitude Score	1	1											
	2	2											
	3	9											
	4	18											
	5	19											
	6	21											
	7	18											
	8	9											
	9	3											
	10	0											
Average Attitude Score		5.5											

Random Sampling Numbers

2017	4228	2317	5966	3861	0210	8610	5155	9252	4425
7449	0449	0304	1033	5370	1154	4863	9460	9449	5738
9470	4931	3867	2342	2965	4088	7871	3718	4864	0657
2215	7815	6984	3252	3254	1512	5402	0137	3837	1293
9329	1218	2730	3055	9187	5057	5851	4936	1253	9640
4504	7797	3614	9945	5925	6985	0383	5187	8556	2237
4491	9949	8939	9460	4849	0677	6472	5926	0851	2557
1623	9102	1996	4759	8965	2784	3092	6337	2624	2366
0450	6504	6565	8242	7051	5501	6147	8883	9934	8237
3270	1772	9361	6626	2471	2277	8833	1778	0892	7349
0364	5907	4295	8139	0641	2081	9234	5190	3908	2142
6249	9000	6786	9348	3183	1907	6768	4903	2747	5203
6100	9586	9836	1403	4888	5107	3340	0686	2276	6857
8903	9049	2874	2104	0996	6045	2203	5280	0179	3381
0172	3385	5240	6007	0671	8927	1429	5524	8579	3196
2756	4979	3434	3222	6053	9117	3326	4470	9314	9970
4905	7448	1055	3525	2428	2022	3566	6634	2635	9123
4974	3725	9726	3394	4223	0128	5958	9269	0366	7382
2026	2243	8808	1985	0812	4765	6563	5607	9785	5679
4887	7796	4339	7693	0879	2218	5455	9375	9726	9077
0872	8746	7573	0011	2707	0520	3085	2221	0467	1913
9597	9862	1727	3142	6471	4622	3275	1932	2099	9485
3799	5731	7040	4655	4612	2432	3674	6920	7210	9593
0579	5837	8533	7518	8871	2344	5428	0048	9623	6645
5585	6342	0079	9122	2901	4139	5140	3665	2611	7832
6728	9625	6836	2472	0385	4924	0569	6486	0819	9121
8586	9478	3259	5182	8643	7384	4560	8957	0687	0815
4010	6009	0588	7844	6313	5825	3711	1847	7562	5221
9455	8948	9080	7780	2689	8744	2374	6620	2019	2652
1163	7777	2320	3362	6219	2903	9415	5637	1409	4716
6400	2604	5455	3857	9462	6840	2604	2425	0361	0120
5094	1323	7841	6058	1060	8846	3021	4598	7096	3689
6698	3796	4413	4505	3459	7585	4897	2719	1785	4851
6691	4283	6077	9091	6090	7962	5766	7228	0870	9603
3358	1218	0207	1940	2129	3945	9042	5884	8543	9567
5249	4016	7240	7305	5090	0204	9824	0530	2725	2088
7498	9399	7830	7947	9692	4558	4037	8976	8441	7468
5026	5430	0188	6957	5445	6988	2321	0569	9344	0532
4946	6189	3379	9684	2834	1935	2873	3959	5634	9707
1965	1344	7839	7388	6203	3600	2596	8676	6790	2168
6417	4767	8759	8140	7261	1400	2828	5586	2338	1615
1843	9737	6897	5656	5795	0188	1189	4807	4260	1192
6558	6087	5109	9661	1553	6681	6688	4475	3701	2888
7990	3100	9114	8565	3175	4315	4593	6478	3453	8802
0723	0015	5905	1609	9442	2040	6376	6567	3411	9410
9008	1424	0151	9546	3032	3319	0014	1928	4051	9269
5382	6202	2182	3413	4103	1285	6530	0097	5630	1548
9817	2615	0450	7625	2033	5484	3931	2333	5964	9627
0891	1244	8240	3062	4550	6454	6517	8925	5944	9995
3721	4677	8487	6739	8554	9737	3341	1174	9050	2962

Each digit is an independent sample from a population in which the digits 0 to 9 are equally likely; that is each has a probability of $\frac{1}{10}$.

ANSWERS

Self-assessed question 1, pp. 44–5.

It is clear that the sampling estimate of the number with a reading problem in the population is wildly inaccurate. It should be 786; in fact it is $10 \times 41 = 410$. One way to avoid the risk of selecting a homogeneous sample of this sort would be to stratify the sample in terms of some known characteristic related to reading attainment, e.g. social class, and select the schools at random from each stratum; i.e. for three social class strata select one school per stratum.

Self-assessed question ■ (b), p. 51.

30.62
30.10
30

Self-assessed question 3, p. 52.

First

Self-assessed question 4, pp. 52–3, Table 5 completed.

Pupil	Score (X)	Deviation (x)	Squared Deviation (x^2)
1	22	7	49
■	12	−3	9
3	14	−1	1
4	12	−3	9
5	14	−1	1
6	14	−1	1
7	17	2	4
8	19	4	16
9	11	−4	16
10	15	0	0
			$\sum x^2 = 106$

$$\text{Mean} = \frac{150}{10} = 15.0$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{106}{10}} = \sqrt{10.6} = 3.26$$

Self-assessed question 5, p. 54.

(a) 95·44
(b) 99·72

Self-assessed question 6, p. 55.

(a) 3·37
(b) restricted
(c) sample homogeneous with respect to attainment in English

Self-assessed question 7, p. 55.

15·9 per cent

Self-assessed question 8, p. 58.

(a) 0·33
(b) $104 \pm 0.66 = 103.34 - 104.66$
(c) $104 \pm 0.85 = 103.15 - 104.85$

References

BUTCHER, H. J. (1966) *Sampling in Educational Research*, Manchester University Press.

FOX, D. J. (1969) *The Research Process in Education*, New York, Holt, Rinehart and Winston.

GUILFORD, J. P. (1965) *Fundamental Statistics in Psychology and Education*, New York, McGraw-Hill (4th edition).

McNEMAR, Q. (1946) 'Opinion-Attitude Methodology', *Psychol. Bull.*, 43, pp. 289–374.

MOSER, C. A. and KALTON, G. (1971) *Survey Methods in Social Investigation*, London, Heinemann.

NISBET, J. D. and ENTWISTLE, N. (1972) *Educational Research in Action*, University of London Press.

PEAKER, G. (1971) *The Plowden Children Four Years Later*, Slough, NFER.

THORNDIKE, R. L. and HAGEN, E. (1969) *Measurement and Evaluation in Psychology and Education*, New York, Wiley (3rd edition).

KERLINGER, F. N. (1969) *Foundations of Behavioural Research*, London, Holt, Rinehart and Winston.

Glossary for Block 2

Terms	*Part and paragraph where it first appears*	*Brief definition*	*See also*
Biased sample	Part 2, 2.15	A sample which fails to reflect accurately the characteristics of the population from which it was drawn.	
Cluster sampling	Part 2, 2.10	A method of sampling entailing random selection of intact groups within a population, e.g. the selection of whole classes or whole schools from a population of school children.	
Confidence interval	Part 2, 5.14	A range of values within which a particular population parameter has a specified probability of lying.	
Deduction	Part 1, 1.7	The process of drawing conclusions from sets of premises by rules of logical inference only.	
Divergent thinking	Part 1, Fig. 2	The production of several solutions to problems which have no single correct answer – e.g. thinking of uses for a milk bottle or listing all the possible strategies for dealing with inflation.	
Dyslexia	Part 1, 2.3	A general term for an inability to understand written or printed verbal symbols.	
Empirical	Part 1, 1.8	Based on experience, i.e. on observation and experiment.	
Extraversion/ introversion	Part 1, 2.3	A personality dimension extensively investigated by H. J. Eysenck measuring the extent to which a person is outgoing, sociable and uninhibited. Extraverts are carefree, spontaneous and impulsive people who like to spend a lot of time in social interaction and whose interests are mainly directed towards the outside world of things and events. At the other extreme are introverts, who are withdrawn, scrupulous, inhibited and serious people, more interested in ideas, imagination and inner life than in social interaction and other exterior events.	
Frequency distribution	Part 2, 5.2	To obtain a frequency distribution the range of values covered by a set of observations is divided into class intervals of equal size, which are set down consecutively in increasing or decreasing order. The number of observations falling in each class interval are then tabulated.	
Frequency Polygon	Part 2, 5.2	A graphical representation of a frequency distribution obtained by plotting the mid-point of a set of class intervals on the X-axis against frequency of occurrence on the Y-axis. The points thus plotted are joined by straight lines.	
Grapheme	Part 1, Activity	*See under* phoneme and grapheme.	
Histogram	Part 2, 5.2	A graphical representation of a frequency distribution. The class intervals are first marked off on the X-axis, and an elongated rectangle is then constructed on each	

		of these marked-off intervals, with a height equal to the number of cases falling in the class interval.	
Homogeneous	Part 2, 2.11	Of the same kind, with respect to the variable under study.	
Hypothesis	Part 1, 1.1	An untested idea or assertion put forward to explain some natural phenomenon. When used in science it has a precise logical form which enables the researcher to predict the outcome of observation and experiment which will confirm or negate the hypothesis.	
Hypothetico-deductive reasoning	Part 1, 1.7	The process of establishing hypotheses by deducing their logical consequents and putting them to empirical test.	*Block 1*
Intelligence	Part 1, 2.3	In its broadest sense intelligence is the capacity to think. Some psychologists, however, limit the term to 'that which intelligence tests measure'. Thus defined, intelligence may be split into several components, such as number, verbal meaning, memory, reasoning and spatial perception.	
Intelligence Quotient (IQ)	Part 2, 1.2	The usual unit of intelligence measurement. For children it used to be defined as 100 times the ratio of mental age to chronological age. General purpose intelligence tests are now usually constructed to yield a mean IQ score of 100 with a standard deviation of about fifteen.	
Junior Eysenck Personality Inventory (JEPI)	Part 1, 3.3	A personality test designed to measure extraversion/introversion and neuroticism/stability in children aged seven to fifteen.	
Level of significance	Part 2, 5.17	*See under* significant.	
Mean	Part 2, 5.2	The main statistic expressing the average value of a set of scores. It is calculated by dividing the total sum of scores Σx by the number of observations N. $$M = \frac{\Sigma x}{N}$$	
Median	Part 2, 5.2	That point above and below which 50 per cent of the values in a distribution lie.	
Minnesota Tests of Creative Thinking (MTCT)	Part 1, Fig. 2	A series of tests purporting to measure creativity in children – involving such tests as imaginative essay writing, improvements to a common object, unusual uses for common objects, and asking all possible questions about a picture and guessing the answers.	
Minor premise	Part 1, 1.9	*See under* major and minor premises.	
Mode	Part 2, 5.2	The most frequent value in a set of observations.	
Moray House Tests	Part 1, Fig. 2	A set of tests measuring various aspects of intelligence, used in the eleven-plus examination.	
Negative relationship	Part 1, 3.4	When two variables vary in such a way that *high* values on the first tend to be associated with a *low* values on the second and vice versa there is said to be a negative relationship between them.	

Neuroticism	Part 1, 3.4	A personality dimension investigated by H. J. Eysenck. The highly neurotic person is apt to be emotionally over-responsive and to breakdown under stress.
Normal distribution	Part 2, 5.6	A bell-shaped mathematical probability distribution relating probability to the value of a variable. It is often closely resembled in nature by the distribution of a population characteristic, e.g. height. Most people cluster round medium height, and as one approaches either extreme the number of cases declines more and more rapidly.
Objective tests	Part 1, 2.6	A test, often of the multiple choice type, where there is a clear and explicit scoring system. Different examiners marking the same person on such a test should arrive at exactly the same score.
Operationalize	Part 1, 3.2	To translate an idea into terms which make it susceptible to objective measurement.
Opportunity sampling	Part 2, 2.14	A non-random method of constructing a sample by selecting cases known to possess certain desired features, without considering for inclusion other cases which possess these features. For example, it might be desired to include in a sample a girls' grammar school of over 1000 pupils. If the choice was based on reasons such as expense or convenience rather than random sampling (e.g. the nearest school was chosen) this would be opportunity sampling.
Parameter	Part 2, 1.2	A descriptive summary of the values of a given characteristic of a population. Examples are the mean, median, mode, range and standard deviation of such values. Cf. statistic. *Note*, this is not the mathematical definition but the definition most commonly adopted in research. (A summary of the values of a given characteristic of a *sample* is a statistic.)
Personality	Part 1, 2.3	A 'blanket' term which has been defined in many different ways. It is commonly taken to refer to the structure and organization of the psychological qualities which characterize a person.
Phoneme and grapheme	Part 1, Activity	Phonemes are the smallest units distinguishable in speech, e.g. the 'd' *sound* in 'dine' or the 'f' *sound* represented by 'gh' in 'cough'. Graphemes are the written letters making up a phoneme e.g. the *letters* 'gh' in 'cough'. Thus grapheme/phoneme relationships are the relationships between the spoken sound and its written representation.
Pilot study	Part 1, 5.1	A small-scale preliminary study undertaken to expose and remedy hidden difficulties and deficiencies in a research design and to develop new instruments.
Population	Part 2, 1.1	Any group of people or observations which includes all possible members in that category, e.g. all six-year-old children in LEA schools in Rugby.
Positive relationship	Part 1, 3.4	When high values on one variable tend to be associated with *high* values on another, and *low* values with *low* values, there is said to be a positive relationship between the two variables.
Positive transfer	Part 1, Activity	When the learning of a previous task facilitates the learning of a subsequent different task, positive transfer is said to occur.

Projective test	Part 1, 2.6	Any test in which the subject in responding to an ambiguous stimulus is supposed to project his own feelings, needs, attitudes, motives etc. into the response. Stimuli used in such tests include inkblots, pictures, incomplete sentences, story themes and single words.
Proportionate stratified sampling	Part 2, 2.6	*See under* stratified sampling.
Quota sampling	Part 2, 2.14	*See* opportunity sampling.
Random sample	Part 1, 4.4	A method of drawing a sample from a population which allows each member of the population an equal chance of being chosen.
Range	Part 2, 5.3	The difference between the largest and smallest values of a set of measurements.
Reliability	Part 1, 2.7	The extent to which a test gives the same result when applied by different people or on different occasions. If a valid estimate of reliability is to be obtained the entity being measured must not alter between applications of the test.
Representative sample	Part 1, 4.4	A sample, drawn from a population, which accurately reflects the characteristics of that population.
Rigidity	Part 1, Fig. 2	An inability to change an inefficient method of problem solving or thinking.
Rorschach test	Part 1, 4.2	A projective personality test in which a subject responds to a standard series of inkblots. The test purports to measure many different personality traits, and there are various scoring systems.
Sample	Part 1, 2.5	A group selected from a population to yield information about this population as a whole.
Sampling distribution	Part 2, 5.10	The distribution of the means of a large number of random samples drawn, with replacement, from the same population.
Sampling error	Part 2, 1.9	No sample, however carefully drawn, can be perfectly representative of the population from which it was drawn. Any estimate of a population parameter made on the basis of a sample is therefore liable to be more or less wrong, depending on the representativeness of the sample. Such errors are known as sampling errors.
Sampling frame	Part 2, 1.6	A list of members of a population from which a sample is to be drawn.
Sampling ratio	Part 2, 2.2	The ratio of the number of people in a sample to the number in the population from which it was drawn.
Sampling with replacement	Part 2, 1.7	Sampling in such a way that when a particular case is chosen the probability of the same case being chosen again is unaffected. For example, such sampling might be done by drawing names from a hat, replacing each name drawn and reshuffling before drawing again.

Sociometry	Part 1, Fig. 2	A technique for measuring the social relationships existing within a particular group. The simplest method used is to ask the people in the group to record privately which members of the group they like, dislike or are indifferent to. Diagrams (sociograms) may then be drawn to portray these relationships.
Stage sampling	Part 2, 2.16	Sampling by first randomly selecting groups from a population, then randomly choosing smaller groups from within these groups and so on.
Standard deviation	Part 2, 5.4	A measure of the spread or dispersion of a set of measurements given by: Standard deviation $= \frac{\Sigma x^2}{N}$ where $\Sigma x^2 =$ the sum of the squares of the deviations from the mean and N is the number of observations.
Standard error	Part 2, 5.12	The standard deviation of a set of means of samples drawn randomly from the same population.
Standardized test	Part 1, 3.3	A test which has been carefully tried out and refined on a large sample of the population for which it is intended, and for which norms are therefore available.
Stratified sampling	Part 2, 1.2	The population to be sampled is first divided into strata, e.g. people of high, medium and low IQ; men and women. A random sample is then drawn from each stratum. Such stratified sampling is said to be proportionate if, in the sample, the proportions of cases from the various strata are the same as those in the population.
Systematic sampling	Part 2, 1.2	The systematic selection of cases from a list at fixed intervals – e.g. every fifth or every tenth case on the list.
Validity	Part 1, 2.7	The extent to which a test actually measures the quality which it was intended to measure.
Variable	Part 1, 3.2	Variables are entities which can take on different values. They may increase or decrease along a given dimension. An independent variable is one which is being systematically varied by the experimenter. A dependent variable is one not directly under the experimenter's control. Its relationship to the independent variable is the object of the investigation.
Verbal reasoning quotient	Part 1, Fig. 2	A quotient similar to intelligence quotient except that it is based only on tests of verbal reasoning.

Notes

Notes

METHODS OF EDUCATIONAL ENQUIRY

Block 1 The Nature of Educational Research

Block 2 Research Design

Part 1 Hypothesis Formulation

Part 2 Sampling

Block 3 Data Collection

Part 1 Practical Aspects of Data Collection

Part 2 Principles of Measurement

Part 3 Principles of Constructing a Measuring Instrument

Part 4 Types of Measure Available

Block 4 Data Analysis

Block 5 Experiment in Educational Research

Block 6 Evaluation and Assessment of Educational Research

KN-712-154

335 03321